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?

The ideal alpha values for Ridge and Lasso are 2 and 0.001, respectively. Using these alpha values, the model achieved an R-squared (R2) score of approximately 0.83.

When we doubled the alpha values for both Ridge and Lasso, the prediction accuracy remained roughly at 0.83. However, there was a slight alteration in the coefficient values. The updated model has been implemented and showcased in the Jupyter notebook.

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?

The optimal lambda values for Ridge and Lasso are as follows:

Ridge: 2

Lasso: 0.0001

The Mean Squared Errors for Ridge and Lasso are as follows:

Ridge: 0.0026203119870136763

Lasso: 0.0024058947123532606

Both models have nearly identical Mean Squared Errors.

Considering that Lasso facilitates feature reduction by setting some feature coefficients to zero, it offers a distinct advantage over Ridge. Therefore, Lasso should be chosen as the final model.

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?

The top five most important predictor variables in the current Lasso model are:

1. GarageArea

2. LotArea

3. TotRmsAbvGrd

4. OverallCond

5. Total_sqr_footage

After removing these attributes from the dataset and building a new Lasso model in the Jupyter notebook, the model's performance changes as follows:

- The R2 (R-squared) of the new model, without the top 5 predictors, decreases to 0.76.

- The Mean Squared Error increases to 0.0036474057020863936

The new top 5 predictors in the revised model are:

1. LotFrontage with a coefficient of 0.199509.

2. HouseStyle_2.5Fin with a coefficient of 0.089089.

3. Total_porch_sf with a coefficient of 0.082716.

4. HouseStyle_2.5Unf with a coefficient of 0.074476.

5. RoofMatl_WdShngl with a coefficient of 0.068923.

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?

According to Occam's Razor, when comparing two models with similar performance on finite training or test data, it is advisable to choose the simpler one. This choice is justified for several reasons:

1. **Generality**: Simpler models tend to be more generic and have broader applicability. They capture the underlying patterns in the data without overfitting to idiosyncrasies present in the training data.

2. **Training Efficiency**: Simpler models require fewer training samples to achieve effective training. They are easier to train and can generalize well even with limited data.

3. **Robustness**: Simple models are often more robust. Complex models can exhibit wild fluctuations in their predictions when exposed to slight changes in the training data. Simple models, on the other hand, are more stable.

4. **Bias-Variance Trade-off**: Occam's Razor highlights the bias-variance trade-off. Complex models have low bias but high variance, meaning they can fit the training data very closely but are prone to overfitting. Simple models have higher bias but lower variance, making them more resistant to overfitting and more reliable on new data.

5. **Generalization**: Simple models generalize better to unseen data because they capture the essential patterns and avoid fitting noise present in the training data.

Regularization techniques can be employed to strike a balance between simplicity and complexity in a model. Regularization adds a penalty term to the cost function, encouraging the model to have smaller parameter values, which effectively simplifies the model.

In summary, maintaining a balance between bias and variance is crucial for model accuracy and generalization. While complex models can fit training data perfectly, they are often too sensitive to changes and do not generalize well. Simpler models might make some errors on the training data but tend to generalize better and are more robust when applied to new data. Occam's Razor reminds us to Favor simplicity in model selection when performance is similar, as it often leads to more reliable and widely applicable models.