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 best alpha value for Ridge regression is found to be 2, and for Lasso regression, it's 0.001. These values were determined to provide a good model fit, with an R2 score of approximately 0.83.

If we were to double these alpha values for both Ridge and Lasso, making them 4 for Ridge and 0.002 for Lasso, the overall prediction accuracy would still be approximately 0.82. However, there would be slight changes in the coefficient values of the predictors in the model.

In simpler terms, even though the model's accuracy remains similar, adjusting the alpha values may causes some minor alterations in how much each predictor variable contributes to the model's predictions.

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 optimum lambda value in case of Ridge and Lasso is as follows:- • Ridge – 2 • Lasso – 0.0001 ▪ The Mean Squared Error in case of Ridge and Lasso are: • Ridge - 0.0018396090787924262 • Lasso - 0.0018634152629407766 ▪ The Mean Squared Error of both the models are almost same. ▪ Since Lasso helps in feature reduction (as the coefficient value of some of the features become zero), Lasso has a better edge over Ridge and should be used 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?

Ans: The five most important predictor variables in the current lasso model is:- 1. Total_sqr_footage 2. GarageArea 3. TotRmsAbvGrd 4. OverallCond 5. LotArea We build a Lasso model in the Jupiter notebook after removing these attributes from the dataset. The R2 of the new model without the top 5 predictors drops to .73 The Mean Squared Error increases to 0.0028575670906482538

To ensure that a model is robust and applicable to a wide range of situations, it's crucial to prioritize simplicity while still maintaining effectiveness. Occam's Razor principle suggests that when two models perform similarly on training or test data, the simpler one should be chosen. This is because simpler models have several advantages:

1. They're more versatile and can be applied to various scenarios.
2. They require fewer training samples to perform well, making them easier and more efficient to train.
3. They tend to be more stable and less prone to drastic changes when the training data is altered.

On the other hand, complex models can suffer from overfitting, where they perform exceptionally well on the training data but poorly on new, unseen data. This is due to high variance, meaning the model is too sensitive to the training data and captures noise instead of true patterns.

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?

To strike a balance between simplicity and effectiveness, regularization techniques can be employed. Regularization helps in keeping the model simple without oversimplifying it to the point of losing usefulness. In regression, regularization involves adding a penalty term to the cost function, which penalizes large parameter values.

Maintaining simplicity in the model also involves understanding the bias-variance trade-off:

- Bias quantifies how accurately the model predicts on new data. A very simple model might have high bias if it fails to capture important patterns in the data.
- Variance measures how much the model's predictions fluctuate with changes in the training data. Complex models tend to have high variance, meaning they're highly sensitive to changes in the training data.

By finding the right balance between bias and variance, we can ensure the model's accuracy is optimized, minimizing errors on both training and test data.