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?

Based on the current model building is :

- optimal value of alpha for ridge model is: 4.0
- optimal value of alpha for Lasso model is: 50

Effect of choosing double the value of optimal alpha:

choosing a value of alpha that is double the optimal value may lead to

- Increased bias
- Decreased variance,
- Slower convergence, and
- Lower accuracy.

It is important to carefully tune the value of alpha to find the optimal balance between bias and variance for the specific dataset and model architecture.

Question 2 : <u>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?</u>

As per Occam's Razor "entities should not be multiplied without necessity" or "the simplest explanation is usually the correct one."

Model complexity depends on two main things:

- No. of features or independent variables and
- Magnitude of beta coefficients.

Normalization (Ridge and Lasso) already shrinks beta coefficients towards zero.

Lasso model is simpler than Ridge with having similar r2 score and MAE.
Ridge:
r2 score on testing dataset: 0.9363995703914922
MSE on testing dataset: 0.018419435953924413
RMSE on testing dataset: 0.135718222630288
MAE on testing dataset: 0.09344961892214859

Lasso:
r2 score on testing dataset: 0.9367982176775831
MSE on testing dataset: 0.01781710976847528
RMSE on testing dataset: 0.13348074680820182
MAE on testing dataset: 0.09142208307508749

As these two models shows almost similar performance on test dataset, so better to choose simpler model, will choose Lasso

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?

Initially top 5 features in Lasso model are as below:

- Constant
- eighborhood_StoneBr
- OverallQual_9
- OverallQual_10
- Neighborhood

After dropping dummy variables, finally

- Neighborhood
- Overall Quality
- Sale condition
- Home functionality
- Overal Condition

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 make model robust and generalisable, better we follow below best practices

Use a diverse and representative dataset:

The dataset used to train the model should be diverse enough to capture the variability in the real-world data. A diverse and representative dataset ensures that the model has learned to recognize patterns and make accurate predictions across different scenarios.

Validate the model on a hold-out dataset:

To ensure that the model is not overfitting to the training data, it should be evaluated on a hold-out dataset that is distinct from the training data.

This evaluation can be done through techniques such as cross-validation or a separate test dataset.

Regularize the model:

Regularization techniques L1/L2 regularization can help prevent overfitting by adding constraints to the model's parameters. Regularization can also help the model generalize better to unseen data.

Monitor the model's performance over time:

The model's performance should be monitored over time to ensure that it remains accurate and generalizable. If the model's performance starts to degrade over time, it may indicate that the model is overfitting to the training data or that the distribution of the input data has changed.

Use transfer learning:

Transfer learning involves using pre-trained models as a starting point for a new task. Transfer learning can help improve the model's performance on a new task with limited training data and also help the model generalize better to new scenarios.