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:

Optimal value for alpha in Ridge Regression is 10

Optimal value for alpha in Lasso Regression is 500

If we double the alpha for ridge :

- Coefficients values are increasing as we doubled alpha.
- r2_score of train data is slightly dropped.

If we double the alpha for lasso :

- Clearly more features are dropped/removed from model when we doubled alpha in lasso.
- r2_score of train data is dropped to ~ 79.6% on train and test data.

Most important predictor variables after the change is implemented :

- Neighborhood_NoRidge , Neighborhood_NridgHt , Neighborhood_Veenker , 2ndFlrSF , Neighborhood_Somerst

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:

I will go ahead with lasso regression as it has further removed 8 variables from total features given by RFE to make it further less complex and generalized. Also the model accuracy is not affected much. It means model will be more simpler and generalized.

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?

<u>Answer</u>:

After excluding the top 5 features, model r2 score has dropped to 75.4% on train set and r2 score has dropped to 76.9% on test set.

5 Top variables now are:

OverallQual , HouseStyle_2Story ,1stFlrSF , KitchenQual , MSZoning_RL

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?

<u>ANSWER:</u>

Ensuring that a machine learning model is robust and generalizable is important to ensure its performance and reliability in real-world scenarios. A robust and generalizable model is less likely to overfit or underfit, and can effectively generalize its predictions to unseen data. Key considerations for achieving model robustness and generalizability:

1. **Sufficient and Representative Data**: The model should be trained on a sufficiently large and diverse dataset that is representative of the real-world data that the model will encounter during deployment. Having enough data helps the model learn patterns that are more likely to be generalizable and less prone to overfitting. If the dataset used for training is too small or biased, the model may not perform well on new data.
2. **Regularization Techniques:** Regularization techniques, such as ridge regression and lasso regression, can help prevent overfitting and improve model robustness. Regularization adds penalties to the model's objective function, which can constrain the model's coefficients and prevent them from becoming overly large or complex. This can lead to a more generalized and robust model.

3. **Cross-Validation**: Cross-validation is a technique used to assess the model's performance on unseen data. It involves dividing the dataset into multiple folds, training the model on some folds and evaluating it on the remaining folds.
4. **Hyperparameter Tuning:** Carefully tuning hyperparameters, such as learning rate, regularization strength, and model complexity, can also impact the model's robustness and generalizability. Grid search or randomized search can be used to systematically search for the optimal hyperparameter values that result in a more robust model.

5. **Model Evaluation**: It is important to evaluate the model's performance using appropriate evaluation metrics that are relevant to the problem domain. Simply relying on accuracy may not always be sufficient, as it may not capture the model's performance on different aspects of the data. It is important to consider metrics such as precision, recall, F1 score, and area under the curve (AUC) that provide a more comprehensive assessment of the model's performance.

The **implications** of model robustness and generalizability for the accuracy of the model are that a robust and generalizable model is likely to have better accuracy on unseen data compared to a model that is not robust or prone to overfitting.

A robust model is less likely to make incorrect predictions or be overly sensitive to changes in the data, leading to improved accuracy. On the other hand, a model that is not robust or generalizable may have lower accuracy, as it may struggle to accurately predict unseen data or may be biased due to overfitting or underfitting. Therefore, ensuring model robustness and generalizability is crucial for obtaining accurate and reliable predictions in real-world scenarios.