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 :

1. Ridge Regression
    a. Before Fine Tuning via GridSearch = 4.0
    b. After Fine Tuning via GridSearch = 4.40
2. Lasso Regression
    a. Before Fine Tuning via GridSearch = 0.001
    b. After Fine Tuning via GridSearch = 0.00059

On doubling the alpha value for Ridge Regression:

1. the Test R2 Score changes from 0.908 to 0.9094
2. the most important predictor variable changes from MSZoning_FV to SaleCondition_Partial

On doubling the alpha value for Lasso Regression:

3. the Test R2 Score changes from 0.9124 to 0.9092
4. the most important predictor variable remains constant = GrLivArea although there is a slight change in the coeff value for the predictor variable.

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:

The choice of optimal value (0.00059) for Lasso Regression should be better as we see that the evaluation metric (namely R2 score) on the test dataset is slightly higher for Lasso Regression. Even with a slight difference, stronger the value, better the regularization.

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:

MSZoning_FV, 2ndFlrSF, Foundation_PConc, 1stFlrSF, GarageAge_New : These would be the new “5 most important predictor variables”, when the Lasso Regression model would encounter a dataset without the current “5 most important predictor variables”.

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:

There are various methods to be carried out to ensure that a model is robust and generalisable. Some of which are as follows:

1. **Regularization:** Carry out regularization techniques such as Ridge and Lasso, to control the complexity of the model and also to prevent overfitting.
2. **Hyperparameter Tuning:** Tuning the hyperparameters using techniques like gridsearch, etc can significantly impact model generalization.
3. **Ensemble Methods:** Making use of methods like Random Forest, which are combinations of multiple models, can lead to increase in predictive performance and generalization of the model.
4. **Outlier Detection and Handling:** Proper detection and handling of outliers is necessary as they can influence the model, causing harm to its generalization capacity.

Increasing model robustness and generalization often involves a trade-off with model accuracy on the training data. As you take steps to prevent overfitting and encourage generalization, the model may not fit the training data as closely. This may lead to slightly lower training accuracy compared to a highly overfit model.
However, the goal is to prioritize the model's ability to make accurate predictions on unseen data (test or validation data) rather than achieving the highest accuracy on the training data.

In summary, robustness and generalization are essential for machine learning models, even if they come at the cost of slightly lower training accuracy. The aim is to strike a balance that results in a model that performs well in real-world scenarios and is less susceptible to overfitting the training data.