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:

The optimal value of alpha in ridge and lasso regression depends on the specific dataset and problem. In the Housing assignment the optimal value of alpha for ridge regression is 3.0 whereas for the lasso it is 0.0001

If we had to double the value of alpha for both ridge and lasso regression, the models would become more regularized and the coefficient estimates would shrink further. This increased regularization would lead to more emphasis on model simplicity and a stronger suppression of less important predictors.

The most important predictor variables after the change is implemented will be 'GrLivArea', '2ndFlrSF', 'BsmtFinSF1' etc.

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 choose Ridge regression as ridge regression is providing the better r2_score value.

Question 3

After building the model, you realized 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:

The five most important predictor variables now are '1stFlrSF', '2ndFlrSF', 'TotalBsmtSF', 'RoofMatl_WdShngl', 'PoolArea'

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:

Here are several key practices to achieve model robustness and generalizability, along with their implications for model accuracy.

a) Sufficient and Representative Data: A robust model requires an adequate amount of high-quality training data that accurately represents the problem domain
b) Train-Test Split: Splitting the data into separate training and testing sets is essential. The training set is used to train the model, while the testing set is reserved for evaluating its performance on unseen data.
c) Cross-Validation: Cross-validation techniques, such as k-fold cross-validation, provide a more robust estimate of the model's performance by repeatedly splitting the data into different train-test folds. This helps assess how the model's performance varies across different data partitions and provides a more reliable estimate of generalization performance.
d) Feature Selection: Careful feature selection contribute to model robustness and generalizability. Irrelevant or redundant features can introduce noise or bias, hindering the model's ability to generalize.
e) Regularization: Regularization techniques, such as ridge and lasso regression, can improve model robustness and generalizability. Regularization introduces a penalty term that discourages complex and overfit models. By controlling the regularization strength through hyperparameter tuning, the model can strike a balance between accuracy and simplicity.
f) Hyperparameter Tuning: Optimizing the model's hyperparameters helps fine-tune its performance.

The implications of ensuring model robustness and generalizability for accuracy lie in achieving a balance between accuracy and the model's ability to generalize to new data. While extreme optimization for robustness and generalizability may slightly impact accuracy on the training data, the overall objective is to build a model that performs well on unseen data and real-world scenarios. This balance ensures that the model can make accurate predictions beyond the training dataset and provides reliable insights in practical settings.