Problem Statement- Part II

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?

Ridge:

when the value of alpha is 2 the test error is minimum.

Lasso:

Initially it came as 0.4 in negative mean absolute error and alpha.when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero.

The most important variable for ridge regression

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

Changes has been implemented for lasso regression

1. GrLivArea
2. OverallQual

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 choice between Ridge and Lasso regression depends on the specific characteristics of the dataset and the goals of your analysis. Here are some factors to consider when deciding between Ridge and Lasso regression:

I will choose Ridge.

Ridge Regression:

- Ridge regression adds a penalty term proportional to the square of the coefficients (L2 regularization).
- It helps to mitigate multicollinearity (high correlation between predictor variables) by shrinking the coefficients towards zero.
- Ridge regression is suitable when you have many predictor variables that are potentially correlated and you want to maintain all of them in the model.
- It can be more effective when there is a small amount of noise in the data.

Lasso Regression:

- Lasso regression adds a penalty term proportional to the absolute value of the coefficients (L1 regularization).
- It encourages sparsity by shrinking some coefficients to exactly zero, effectively performing feature selection.
- Lasso regression is suitable when you have a high-dimensional dataset with many irrelevant or redundant features, and you want to identify the most important predictors.
- It can be more effective when there are only a few predictor variables that have significant effects, while others have little to no effect.

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?

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageArea

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?

High-quality training data: The model should be trained on a diverse and representative dataset that covers a wide range of scenarios and examples. The data should be clean, well-labeled, and free from biases that could lead to skewed results.

Regularization techniques: Regularization methods like L1 and L2 regularization, dropout, and early stopping can prevent overfitting by adding penalties or introducing randomness during training. Regularization encourages the model to focus on important features and reduces reliance on noise or irrelevant patterns.

Testing on unseen data: Evaluating the model on a separate test set that was not used during training provides an estimate of its performance on new, unseen data. This step helps assess generalization capabilities and detect potential issues.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data