Problem Statement – Part 2

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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans

In ridge regression, as we increase the alpha value from 0, the mean absolute error decreases, but the train error shows an increasing trend. At alpha = 2, the test error is minimized, so we choose alpha = 2 for ridge regression.

For lasso regression, I opt for a small alpha value of 0.01. Increasing alpha leads the model to penalize more, making most coefficient values approach zero. Initially, the negative mean absolute error and alpha are 0.4.

Doubling the alpha value to 10 in ridge regression intensifies the penalty, aiming for a more generalized model, simplifying it without overfitting. However, both test and train errors increase significantly.

Similarly, raising alpha in lasso regression increases penalization, pushing more coefficients towards zero, resulting in a decrease in the R2 square value.

Following changes in ridge regression, the most crucial variables are:

- MSZoning_FV
- MSZoning_RL
- Neighborhood_Crawfor
- MSZoning_RH
- MSZoning_RM
- SaleCondition_Partial
- Neighborhood_StoneBr
- GrLivArea
- SaleCondition_Normal
- Exterior1st_BrkFace

After implementing changes in lasso regression, the most important variables are:

- GrLivArea

- OverallQual
- OverallCond
- TotalBsmtSF
- BsmtFinSF1
- GarageArea
- Fireplaces
- LotArea
- LotFrontage

<u>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>Ans</u>

Regularizing coefficients is crucial to enhance prediction accuracy while reducing variance and maintaining model interpretability. In ridge regression, a tuning parameter, lambda, is employed to penalize the square of coefficient magnitudes, determined through cross-validation.

The penalty, lambda multiplied by the sum of squared coefficients, targets variables with larger values, mitigating their impact. Increasing lambda diminishes model variance while keeping bias constant, and ridge regression incorporates all variables in the final model.

Lasso regression also employs a tuning parameter, lambda, penalizing the absolute value of coefficient magnitudes, identified through cross-validation. As lambda increases, Lasso progressively shrinks coefficients towards zero, effectively setting some variables to exactly zero. Lasso conducts variable selection, and when lambda is small, it performs like simple linear regression.

With increasing lambda, shrinkage occurs, leading the model to neglect variables with zero values.

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?

Ans

The following five crucial predictor variables are set to be omitted:

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- GarageArea

Question 4
How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Ans

The goal is to keep the model simple for robustness and generalizability, even if it means sacrificing some accuracy. This concept aligns with the Bias-Variance trade-off. A simpler model tends to have higher bias but lower variance, making it more generalizable. In practical terms, a robust and generalizable model should exhibit consistent performance on both training and test data, maintaining accuracy across different datasets.

Bias: This represents the error in a model when it struggles to learn from the data. High bias implies the model cannot grasp the details in the data, resulting in poor performance on both training and testing data.

Variance: On the other hand, variance signifies the error in a model when it attempts to over-learn from the data. High variance indicates the model excels on training data but falters on testing data, as it may not have seen that particular data before.

Balancing bias and variance is crucial to avoid overfitting (capturing noise in the data) and underfitting (oversimplifying the data). Achieving this balance ensures the model performs well across various datasets.