Problem Statement - Part II

Assignment 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?

Answer:
In Ridge Regression, as we vary alpha and plot the curve between negative mean absolute error and alpha, we observe a decreasing trend in the error term as alpha increases from 0. However, the train error exhibits an increasing trend with higher values of alpha. At an alpha value of 2, the test error is minimized, leading us to choose alpha equal to 2 for our ridge regression.

Regarding Lasso regression, I've opted for a very small alpha value of 0.01. As we increment alpha, the model intensifies its effort to penalize coefficients, aiming to force many coefficients to zero. Initially, the negative mean absolute error stands at 0.4 when alpha is increased. Doubling the alpha value to 10 in our ridge regression intensifies the penalty on the curve, seeking to simplify the model and avoid overfitting every data point in the dataset. However, with alpha set to 10, we observe increased error for both the test and train datasets.

Similarly, increasing alpha for Lasso results in a more stringent penalty on our model, leading more coefficients to be reduced to zero. As alpha rises, the R-squared value decreases.

After implementing changes, the most significant variables for Ridge Regression are:

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

For Lasso Regression, the most important variables are:

- GrLivArea
- Overall Qual
- OverallCond
- TotalBsmtSF
- BsmtFinSF1
- GarageArea
- Fireplaces
- LotArea
- LotFrontage

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:
Ensuring the regularization of coefficients is crucial for enhancing prediction accuracy while simultaneously reducing variance and enhancing model interpretability. Ridge regression introduces a regularization parameter, often denoted as lambda, to penalize the square of the coefficients. This penalty term, determined via cross-validation, aims to minimize the residual sum of squares, effectively shrinking coefficients with larger magnitudes. As lambda increases, the variance of the model decreases while the bias remains constant, ultimately including all variables in the final model. Conversely, Lasso regression utilizes a similar lambda parameter but penalizes the absolute value of the coefficients. Through cross-validation, Lasso progressively shrinks coefficients towards zero, potentially resulting in some variables being exactly zero, thus performing variable selection. Lower lambda values yield a model akin to simple linear regression, while increasing lambda leads to shrinkage and eventual elimination of variables with zero coefficients.

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:
The five most crucial predictor variables that will be excluded are as follows:

- Ground Living Area (GrLivArea)
- Overall Quality (Overall Qual)
- Overall Condition (Overall Cond)
- Total Basement Area (TotalBsmtSF)
- Garage Area (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?

Answer:
Simplifying the model is imperative to enhance its robustness and generalizability, even at the expense of reduced accuracy. This principle aligns with the Bias-Variance trade-off, where a simpler model typically exhibits higher bias but lower variance, resulting in greater generalizability. This trade-off implies that a robust and generalizable model performs consistently well on both training and test data, with minimal deviation in accuracy between the two datasets.

Bias, in this context, represents the error inherent in the model when it struggles to learn from the data. High bias indicates that the model fails to capture intricate patterns within the data, leading to poor performance on both training and testing datasets.