Abhi - Problem Statement - Part II

Q1. 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?

A: In ridge regression, as we increase the alpha value from 0, we observe a decreasing trend in the error term. However, the training error increases with higher alpha values. When alpha reaches 2, we achieve the minimum test error, leading us to select alpha = 2 for ridge regression.
For lasso regression, I've opted for a small alpha value of 0.01. Increasing alpha intensifies the penalty on coefficients, tending to shrink them towards zero. Initially, the negative mean absolute error and alpha stood at 0.4. Doubling alpha to 10 in ridge regression leads to a more substantial penalty, making the model more general and simpler, thereby not overfitting the data. Unfortunately, this increase in alpha results in higher errors for both training and testing data.

Similarly, when we elevate alpha in lasso, the model penalizes coefficients further, forcing more of them to become zero. As alpha increases, the R-squared value decreases.

The most important variable after the changes has been implemented for ridge regression are as follows:-

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

The most important variable after the changes has been implemented for lasso regression are as follows:-

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage

Q2. 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?

A: Regularizing coefficients is essential for enhancing prediction accuracy, reducing variance, and promoting model interpretability. In Ridge regression, lambda, a tuning parameter identified through cross-validation, serves as the penalty, which is the square of the coefficient magnitudes. The penalty term is lambda times the sum of squared coefficients, penalizing those with higher values. As lambda increases, the model's variance decreases while bias remains constant. Unlike Lasso regression, Ridge regression retains all variables in the final model.

Lasso regression also employs a tuning parameter, lambda, serving as the penalty, which is the absolute value of the coefficient magnitudes determined through cross-validation. As lambda increases in Lasso, it drives coefficients towards zero, ultimately setting some variables exactly to zero, effectively performing variable selection. When lambda is small, Lasso behaves like a simple linear regression, and as lambda increases, it performs variable shrinkage, neglecting variables with a coefficient value of zero.

Q3. 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?

A: Those 5 most important predictor variables that will be excluded are :-

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

Q4. 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?

A: Simplicity in a model is key, even though it may result in decreased accuracy. However, this simplicity leads to a more robust and generalizable model, which can be explained through the Bias-Variance trade-off. In essence, a simpler model exhibits higher bias but lower variance, making it more generalizable. The implication for accuracy is that a robust and generalizable model performs consistently well on both training and test data, with minimal variations in accuracy between them.
Bias: Bias represents errors in the model's ability to learn from the data. High bias indicates that the model struggles to capture the finer details in the data, resulting in poor performance on both training and testing data.
Variance: Variance signifies errors in the model when it attempts to excessively learn from the data. High variance means that the model excels on the training data it has seen extensively but fares poorly on testing data, which it has not encountered before.
Achieving a balance between Bias and Variance is crucial to avoid both overfitting and underfitting of data.