Problem Statement - Part II

Q1. 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?

Ans: Regularizing coefficients is crucial for enhancing prediction accuracy while simultaneously reducing variance and maintaining interpretability in the model.

Ridge regression employs a regularization parameter known as lambda to penalize the square of the coefficient magnitudes, which is determined through cross-validation. By imposing a penalty proportional to lambda times the sum of squared coefficients, Ridge regression penalizes coefficients with higher values, ultimately minimizing the residual sum of squares. As lambda increases, the variance decreases while the bias remains constant. Unlike Lasso Regression, Ridge regression retains all variables in the final model.

On the other hand, Lasso regression utilizes lambda as the penalty to shrink the coefficients toward zero, with the penalty being the absolute value of the coefficient magnitudes determined via cross-validation. Increasing lambda in Lasso regression leads to the coefficients approaching zero, effectively setting some variables to exactly zero. Lasso regression also performs variable selection: for small lambda values, it behaves similarly to simple linear regression, but as lambda increases, shrinkage occurs, resulting in certain variables being disregarded by the model.

Q2. 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?

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

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

Q3. 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: Simplicity is key for the model, even though it may result in decreased accuracy, it enhances its robustness and generalizability. This concept can be grasped through the Bias-Variance trade-off. A simpler model tends to exhibit higher bias but lower variance, making it more generalizable. The implication for accuracy is that a robust and generalizable model should perform consistently well on both training and test data, showing minimal change in accuracy between the two datasets.

Bias refers to the error in the model when it lacks the capacity to effectively learn from the data. High bias indicates that the model struggles to capture the nuances within the data, resulting in poor

performance on both training and testing datasets.

Variance, on the other hand, represents the error in the model when it overly learns from the data. High variance suggests that the model performs exceptionally well on the training data but poorly on the testing data, as it struggles to generalize to unseen data.

Maintaining a balance between bias and variance is crucial to avoid overfitting or underfitting of the data.

Q4. 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?

Ans: For ridge regression, as alpha increases from 0, the negative mean absolute error decreases, but the train error exhibits an increasing trend with higher alpha values. When alpha equals 2, the test error is minimized, leading us to select alpha equal to 2 for ridge regression.

In the case of lasso regression, I've opted for a very small alpha value of 0.01. With increasing alpha, the model imposes more penalties, aiming to drive most coefficient values towards zero. Initially, the negative mean absolute error was 0.4 at alpha.

Doubling the alpha value for ridge regression to 10 increases the penalty applied to the model, aiming to generalize it further by simplifying the model and reducing the need to fit every data point in the dataset. However, from the graph, it's evident that with an alpha of 10, both test and train errors increase.

Similarly, increasing alpha for lasso regression results in further penalization of the model, leading more coefficients to be reduced to zero. As alpha increases, the R-squared value also decreases.

The most significant variables after implementing these changes for ridge regression are as follows:

1. MSZoning_RL
2. MSZoning_FV
3. MSZoning_RH
4. MSZoning_RM
5. Neighborhood_Crawfor
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. OverallCond
2. GrLivArea
3. LotArea
4. Fireplaces
5. GarageArea

6. OverallQual
7. LotFrontage
8. BsmtFinSF1
9. TotalBsmtSF