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, reducing variance, and ensuring model interpretability.

Ridge regression employs a regularization parameter, lambda, determined through cross-validation. The penalty in Ridge regression is the square of the magnitude of coefficients, aiming to minimize the residual sum of squares. By introducing a penalty term, lambda times the sum of squared coefficients, the model penalizes larger coefficients, effectively reducing variance while maintaining a constant bias. Ridge regression includes all variables in the final model, distinguishing it from Lasso Regression.

In Lasso regression, lambda serves as the tuning parameter, with the penalty being the absolute value of the coefficients' magnitude, determined through cross-validation. As lambda increases, Lasso systematically shrinks coefficients towards zero, potentially setting some variables exactly to zero. This property makes Lasso regression not only a regularization method but also a variable selection technique. With a small lambda, Lasso performs similarly to simple linear regression, but as lambda increases, the model exhibits shrinkage, ultimately neglecting variables by assigning them zero coefficients.

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: Achieving a robust and generalizable model involves diverse dataset usage, cross-validation, and regularization techniques. Robustness ensures adaptability to variations, while generalization ensures performance on unseen data. Balancing both aspects is crucial for high accuracy across diverse scenarios. Overfitting is reduced, allowing the model to excel in real-world applications. The interplay of robustness and generalization optimizes accuracy and reliability in various contexts. The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalizable. It can be also understood using the Bias-Variance trade-off.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

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.

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: In the case of ridge regression: - When we plot the curve between negative mean absolute error and alpha, we see that as the value of alpha increase from 0 the error term decrease and the train error is showing increasing trend when value of alpha increases. when the value of alpha is 2 the test error is minimum so we decided to go with value of alpha equal to 2 for our ridge regression.

For lasso regression I have decided to keep very small value that is 0.01, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero. Initially it came as 0.4 in negative mean absolute error and alpha.

When we double the value of alpha for our ridge regression no, we will take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more generalized that is making model simpler and no thinking to fit every data of the data set. from the graph we can see that when alpha is 10, we get more error for both test and train.
Similarly, when we increase the value of alpha for lasso, we try to penalize more our model and more coefficient of the variable will be reduced to zero, when we increase the value of our r2 square also 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