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?

Ans.) In ridge regression, we observed that as the value of alpha increases, the negative mean absolute error decreases, but the train error shows an increasing trend. At alpha value 2, the test error is minimized, so we chose alpha equal to 2 for ridge regression. For lasso regression, I decided to use a small value of 0.01 for alpha. Increasing the value of alpha in lasso regression leads to greater penalization and more coefficients approaching zero. Initially, the negative mean absolute error for lasso regression was 0.4 at alpha value 0.01. Doubling the alpha value in ridge regression to 10 increases the penalty and aims to simplify the model by making it more generalized, resulting in higher errors for both test and train data. Similarly, increasing alpha in lasso regression penalizes the model more, reducing more variable coefficients to zero and decreasing the value of r2 square.

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

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

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

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?

Ans.) Regularizing coefficients is crucial for improving prediction accuracy, reducing variance, and enhancing model interpretability. In ridge regression, a tuning parameter called lambda is utilized as the penalty, which is determined through cross-validation. The objective is to minimize the residual sum of squares by applying the penalty, where the penalty itself is lambda multiplied by the sum of squares of the coefficients. Consequently, coefficients with larger values are penalized more heavily. By increasing lambda, the variance in the model decreases while the bias remains constant. Unlike Lasso regression, ridge regression includes all variables in the final model.

In Lasso regression, a tuning parameter called lambda is also employed as the penalty, determined through cross-validation. As the lambda value increases, Lasso regression shrinks the coefficients towards zero, effectively setting some variables exactly equal to zero. This process facilitates variable selection. When the lambda value is small, Lasso regression performs similar to simple linear regression. However, as the lambda value increases, shrinkage occurs, and variables with a coefficient of zero are disregarded by the model.

Q3.) 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.) Those 5 most importantpredictorvariables 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 generalisable? What are the implications of the same for the accuracy of the model and why?

Ans.) Simplicity is a key aspect for a model, even if it leads to a decrease in accuracy, as it enhances robustness and generalizability. This concept can be understood through the Bias-Variance trade-off. A simpler model exhibits higher bias but lower variance, making it more generalizable. From an accuracy perspective, a robust and generalizable model performs equally well on both training and test data, with minimal changes in accuracy.

Bias refers to the error in a model when it is unable to effectively learn from the data, indicating a weak learning capacity. High bias implies that the model fails to capture intricate details within the data, resulting in poor performance on both training and testing datasets.

Variance, on the other hand, represents the error in a model when it tends to over-learn from the data. A high variance indicates that the model performs exceptionally well on the training data since it has memorized the specific patterns within it. However, when faced with unseen testing data, the model's performance deteriorates significantly due to its inability to generalize. Achieving a balance between bias and variance is crucial to avoid both overfitting and underfitting of the data, ultimately leading to optimal model performance.