PROBLEM STATEMENT-PART 2

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?**

Answer:- In the case of ridge Regression,when we plot the curve between r2 score and alpha,we see that as the value of alpha increase from zero,the error term decrease and the train error is showing increasing trend when the value of alpha increases. When the value of alpha is 0.9,the test error is minimum,so we decide to go with value of alpha equal to 0.9 for our ridge regression.

For Lasso regression,I have used smaller value of alpha as 0.001 and when we increase the value of alpha,the model will penalize more and more and try to make most of the coefficient value equal to zero.
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 more 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 our model more and coefficients of the variable will reduced to zero, when we increase the value of our r2 square also decreases.

The most important predictor variables after the change is implemented for lasso regression are as follows:-

1. MiscVal
2. BsmtHalfBath
3. LowQualFinSF
4. BsmtFullBath
5. HalfBath
6. BsmtFin Type1
7. Neighborhood_Gilbert
8. LotShape
9. HeatingQC
10. Neighborhood_BrkSide

The most important predictor variables after the change is implemented for Ridge regression are as follows:-

1. MiscVal
2. BsmtHalfBath
3. HalfBath
4. LowQualFinSF
5. BsmtFullBath
6. Neighborhood_Gilbert

7. EnclosedPorch
8. TotRmsAbvGrd
9. GrLivArea
10. Neighborhood_IDOTRR

Q 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:- It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model.
Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant.
Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

After creating models in both Ridge and Lasso,we can see that r2_scores are almost same for both of them but as lasso will penalize more on dataset and help in feature elimination,I will consider lasso as my best 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?**

Answer:- The five most important predictor variables are:-
1.BsmtFinType1
2.Neighborhood_Gilbert
3.LotShape
4.HeatingQC
5.Neighborhood_BrkSide

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?**

Answer:- The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less

variance and more generalisable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.
Bias: Bias is the 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 over-fitting and under-fitting of data.