Problem Statement - Part II

Question 1

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: The value for alpha ridge is equal to 2: When we plot the curve between negative mean absolute error and alpha we see that the value of alpha increases from zero, the error term decreases and the train error is showing increasing trend wen 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.

Lasso regression: The value is 0.01, when we increase the value of alpha the model tries to penalize more and try to make most of the coefficient value to zero. Initially it came as 0.4 in negative mean absolute error and alpha.

Changes in the model if you choose to double the value of alpha for both ridge and lasso: The value of alpha for ridge regression 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 will make model simpler. When alpha is 10 we get more error for both test and train data set.

The value of alpha for lasso will try to penalize our model more and coefficient of the variable will reduce to zero, when we increase the value and r2 square also decrease.

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. **GarageArea**
6. **Fireplaces**
7. **LotArea**
8. **LotFrontage**
9. **BsmtFinSF1**
10. **BsmtFullBath**

Question 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?

Ans: Ridge regression uses a tunning parameter called Lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values get penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant, ridge regression includes all the variables in final model unlike Lasso regression.

Lasso regression uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficient which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it makes the variables exactly equal to zero. Lass 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

Moreover, while choosing a type of regression in the real world an analyst must deal with the lurking and confounding dangers of outliers, overfitting. Using ridge results in exposing the analyst to such risk. Hence use of Lasso could be quite beneficial as it is quite robust to fend off such risks to a large extent, thereby resulting in better and robust regression models.

Question 3

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: The five most important predictor variables are:

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

Question 4

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: A model is robust when any variation in the data does not affect its performance much.

A generalizable model can adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.

To make sure a model is robust and generalizable, we must take care it doesn't overfit. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data but fail to pick up the patterns in unseen test data.

In other words, the mode should not be too complex to be robust and generalizable

Too much importance should not be given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. If we look at it, the perspective of accuracy, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. The addition of bias means that accuracy will decrease.

In general, we must strike some balance between model accuracy and complexity. This can be achieved by regularization techniques like Ridge regression and Lasso.