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: Regularize coefficients is important and it will improve the prediction accuracy and making the model good.

Ridge regression uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model.

As 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.

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?

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:. The simpler the model the more the bias but less variance and more generalizable. The 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: High bias model isunable to learn details in the data. Model performs poor on training and testing data.

Variance: High variance meansmodel performs exceptionally well on training data.
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: Ridge regression:- When we plot the curve between negative mean absolute error and alpha the value of alpha increase from 0 the error term decrease and the train erroris showing increasing trend when value of alpha increases .

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.

1. Neighborhood_Crawfor
2. MSZoning_RH
3. MSZoning_RM
4. SaleCondition_Partial
5. Neighborhood_StoneBr
6. GrLivArea
7. SaleCondition_Normal
8. Exterior1st_BrkFace
9. MSZoning_RL
10. MSZoning_RM
11. GrLivArea
12. MSZoning_FV

The important variable after the changes been made for lasso regression are:-

1. OverallCond
2. TotalBsmtSF
3. GarageArea
4. Fireplaces
5. BsmtFinSF1
6. LotArea
7. LotArea
8. GrLivArea
9. OverallQual
10. LotFrontage