Machine Learning 2 Assignment 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?

Answer 1:

A curve between the mean error and alpha is plotted to find the optimal value for alpha. For ridge regression the optimal value is 2. For lass the optimal value is 0.01 in our case.

When we double the values for alpha, the model will become more generalizable and simpler trying to fit more data and the error will increase. In lasso aswell more coefficients will reduce to zero.

Important variables for ridge after the change:

1. GrLivArea
2. OverallQual
3. OverallCond
4. LotArea
5. BsmtFinSF1

Important variables for Lasso after change:

1. OverallQual
2. GrLivArea
3. GarageArea
4. TotalBSmntSF
5. Fireplaces

Question 2:

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now which will you chose to apply and why?

Answer 2:

Ridge regression gives good results in terms of R2 values of test and train data sets, Lasso regression would be a preferred choice since it assigns a zero value to insignificant variables enabling to choose proper predictor variables. It is better to use simple and robust models than complicated ones with multiple features.

Question 3

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 have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer 3:

The five most important predictor variables that need to excluded are:

1. PoolQC
2. Neighbourhood
3. Exterior1st_BrkFace
4. Age of House
5. SaleCondition

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?

Answer 4:

Generalization refers to the model's ability to adapt to new and previously unseen data drawn from the same distribution as the one used to create the model. The model can be made generalizable by developing training environments or samples of data, where we can maximize the accuracy while also guaranteeing predictor invariance. The predictors that both fit the data well and are invariant across environments are used as outputs in the final model.

We would require that our training set has good variance for better model building. K-fold cross validation can be used for training and testing on k different subsets of the data.

Hyperparameter tuning using regularization also helps in generalizing for better performance.