# CPSC-6300 **APPLIED DATA SCIENCE**

## ANALYSIS OF ENERGY EFFICIENCY DATA SET

Chundru Rohith Venkata Sai Ram

**INTRODUCTION**: In this project, I am going to analyze how various model like Multiple Linear Regression, k-Fold, LOOCV, Regression Trees, Random Forests, Bagging and Boosting performs for the energy efficiency dataset for different houses. Model performance is determined by finding the mean square error of the model by using model predicted values and actual values of the responses. The data contains about the various factors of the houses which effects the heating load(Y1) and cooling load(Y2) of the houses.

**DATASET:** The Dataset consists of 1296 observations with 10 variables of which 8 are predictor variables and 2 are response variables. The Data Set is taken from UCI Machine learning Repository where it contains information about factors that are responsible for energy efficiency of the Houses. The following 8 features(X1,X2,X3,X4,X5,X6,X7,X8) are useful to predict the both of the responses(Y1,Y2) .All these attributes are real values.

Features Information of the Houses:

X1-> Relative Compactness X2-> Surface Area X3->Wall Area

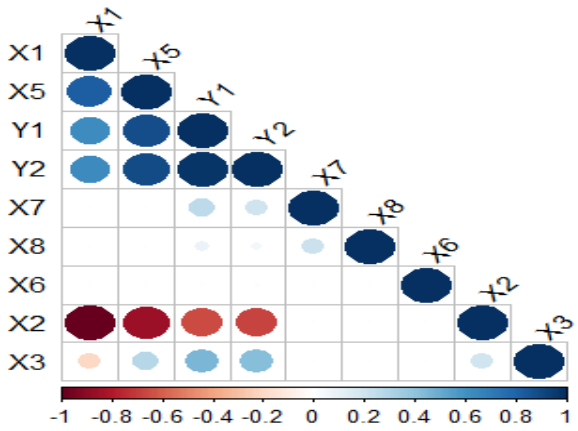
X4-> Roof Area X5-> Overall Height X6-> Orientation

X7->Glazing Area X8-> Glazing Area Distribution

Y1-> Heating Load (Response1) Y2->Cooling Load (Response2)

URL for dataset: https://archive.ics.uci.edu/ml/datasets/Energy+efficiency#

The Dataset contains 2 full columns and few rows contain NA values. I removed those NA values by using na.omit() function. For selecting best predictors, used forward and backward selection for both responses Y1 and Y2 and output of these selections removed X4. Correlation Matrix after removing X4 is shown below.




**Model Selection and validation:** The models selected for analyzing this dataset are Multiple Linear Regression, K-Fold , LOOCV, Fitting Regression Trees, Bagging, Random Forest and Boosting.

**1)Multiple linear regression**: For multiple linear regression, Mean Square Error for Y1 as response is 10.13 and Y2 as response, I got Means Square Error of 11.26187.

**2)K-Fold**: Response as Y1: By taking K as 10, implemented K -Fold using cv.glm() function got the cv.errory1.10$delta as 101.89 101.88 which represents raw cross validation estimate and adjusted cross validation estimate.

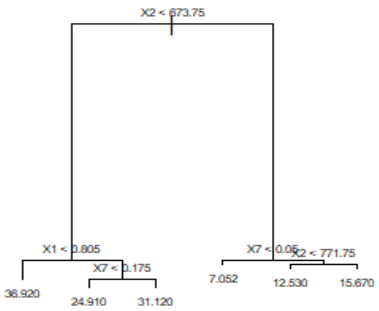
Response as Y2: By taking K as 10, got the cv.errory2.10$delta as 10.3,10.3

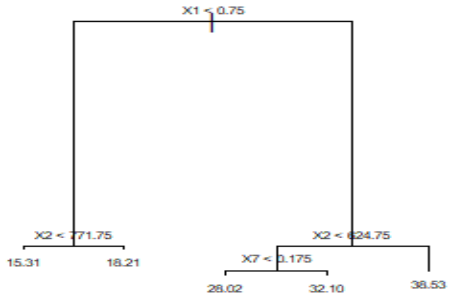
**3)LOOCV**: Response as Y1:Got the loocv.errory1$delta as 101.94 101.94

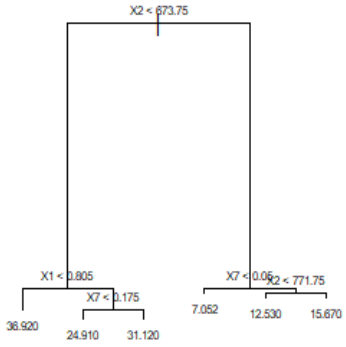
Response as Y2:Got the loocv.errory2$delta as 10.34 10.34

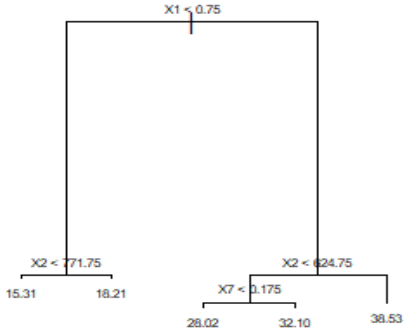
**Splitting data**: For all tree based dataset is divided equally into  train set and test set where each model  implementation is done for both responses Y1 and Y2

**5)Fitting Regression Trees**: Firstly used tree() function for fitting the train data and plotted that tree with text. Secondly, by using cv.tree() function to check whether pruning will improving performance and pruned the tree using prune.tree() function. *Left side plots are for response variable Y1 and right side plots are for response variable Y2.*



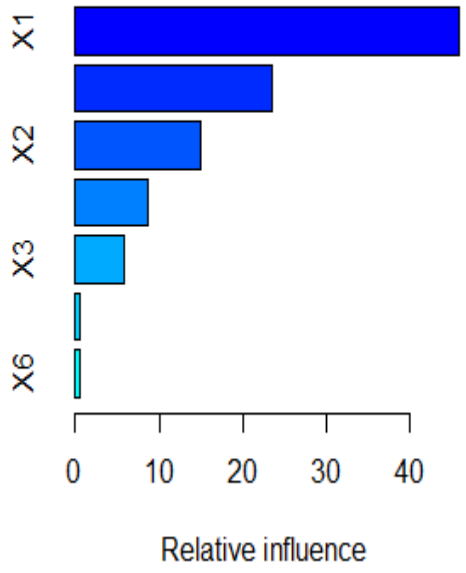



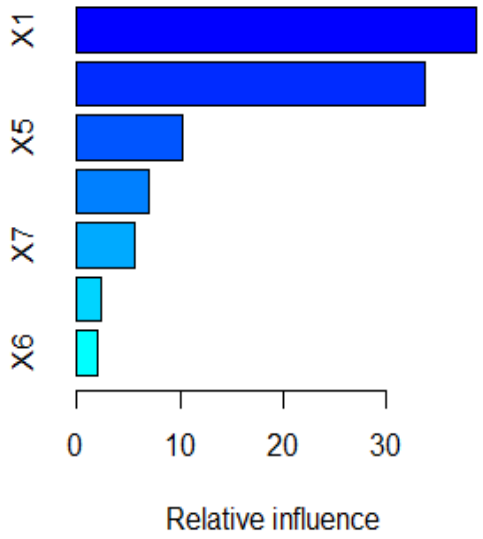
*Top left* and *top right* trees are for *fitting regression trees* of two response variables Y1 and Y2 whereas *bottom left* and *bottom right* are *pruned trees* of two response variables Y1 and Y2. The Mean Square Error for this model is nearly 9.37 for Y1 and around 9.71 for Y2.

**6)Bagging**: Bagging is done by loading the library of randomforest and by taking the no. of variables randomly sampled(mtry) = 7 since 2 are response variables already. Mean Square Error obtained by bagging model for Y1 and Y2 is 0.57 and 3.5.

**7)RandomForest:** Random Forest is also done like bagging where mtry is taken as 4 where as default mtry should be mtry/3 for random forest. Mean Square Error obtained by random forest model is 0.82 for Y1 and 3.4 for Y2.

**8)Boosting:** Gbm library is loaded initially for Boosting and implemented using gbm function with no. of trees =1000 ,distribution = gaussian. It used the training data like other tree models and calculated the error for test dataset which is obtained as 0.35 for Y1 and 1.96 for Y2.



By looking the summary of Boosted Regression Model we got the plot of predictor variables relative influence with respect to response Y1(left plot) and Y2(right plot).

**CONCLUSION:** By observing Mean Squared Errors from the implemented models. Linear regression performs badly when compared to tree based models because linear regression has an MSE of 10.13 for Y1 and 11.26 for Y2 whereas for regression tree it is 9.37 and 9.71 which is the least performer of all the tree based models. The order for best performed models to this dataset is Boosting > Bagging > Random Forest > fitting regression tress > multiple linear model.