CS 40003 : Data Analytics

Project Assignment #3 26.08.2016

You have given a data set concerning housing values in suburbs of Mumbai, India. The following is the meta-data of the data set.

Attribute	Attribute's information
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft
INDUS	Proportion of non-retail business acres per town
CHAS	Mahim River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five industry employment centres
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per Rs. 5,00,000
PTRATIO	Pupil-teacher ratio by town
MINO	$1000\times(Minor - 0.63)^2$ where *Minor* is the proportion of minorities by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in Rs. 5,00,000's

A data sheet in Microsoft Excell includes 506 sample data. Use the data set and perform the following.

1. Partition the data into training and test sets.
2. Use training set to fit a multiple linear regression model to the median house price (MEDV) as response variable and other variables as repressors.
3. Compute test R^2 (using test data set).
4. Compute the correlation table for the 13 numerical predictors and choose which one to remove based on this table.
5. Try to reduce test R^2 by choosing the right set of repressor variables.

Submission procedure:

1. Submit R code and a report describing the process you have followed and results therein. Submit your file to Moodle course management system.
2. Each group should submit only one submission of the above.
3. Last date of submission is: **06.11.2016, 23:55 hours**.