Capstone Project - The Battle of the Neighborhoods – Predicting House Value
Coursera / Applied Data Science Capstone by IBM Foursquare

Note for the Peer Reviewer:

Because the Foursquare developer account is currently unavailable in the country I'm located (even though I contacted the Foursquare company), I can't access the Foursquare location data for this project.

Thus I have to use other kinds of data and technologies for the project instead.

Data

2.1 Data sources

Boston house prices dataset Dataset derived from information collected by the U.S. Census Service concerning housing in the area of Boston Mass.

The dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. "It was obtained from the StatLib archive (http://lib.stat.cmu.edu/datasets/boston), and has been used extensively throughout the literature to benchmark algorithms". The dataset has 506 cases.

There are 14 attributes in each case of the dataset. They are:

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 - Charles 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 Boston employment centres
RAD - index of accessibility to radial highways
TAX - full-value property-tax rate per $10,000
PTRATIO - pupil-teacher ratio by town
B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT - % lower status of the population
MEDV - Median value of owner-occupied homes in $1000's

sklearn.datasets of python also contains the dataset, which is used for this project.

2.2 Data preparation

Data are loaded from sklearn.datasets of python 3.
The dataset obtained has already been cleaned and thus contains no null data or missing values.
The dataset is then divided into training dataset and test dataset, with 70% : 30% in proportion.
MEDV - Median value of owner-occupied homes in $1000's is defined as the target variable.

2.3 Feature selection

There is 506 samples and 13 features in the data. Feature selection analysis is done by using the feature_importances_ *property of* GradientBoostingRegressor. Although the feature_importances_ of 2 variables (“ZN - proportion of residential land zoned for lots over 25,000 sq.ft.” and “ CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)”) are close to 0.0001, intuitively all the variables are not related in nature. Additionally, there are only 13 features, which are quite limited. Therefore, 13 features are selected.