Capstone Project - Predicting House Value

Coursera / Applied Data Science Capstone by IBM Foursquare

Introduction: Business Problem

1.1 Background

House prices affect the life of most people. Therefore, it is beneficial to accurately predict house prices. Many factors can cause house prices to rise or fall. In order to accurately predict house prices, data has to be collected, impacting factors need to be determined, and appropriate models should be developed.

In this project, Boston house prices dataset is analyzed with machine learning algorithms to predict the value of houses.

1.2 Problem

This project aims to predict the value of houses. Factors (features) that might contribute to determining the value of houses might include per capita crime rate by town, proportion of residential land zoned, nitric oxides concentration, average number of rooms per dwelling, index of accessibility to radial highways, etc. Feature selection need to be conducted, and predicting models are to be developed and evaluated.

1.3 Interest

Buyers and sellers of houses, whether ordinary residents or organizations, would be interested in predicting the value of houses.

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.

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.

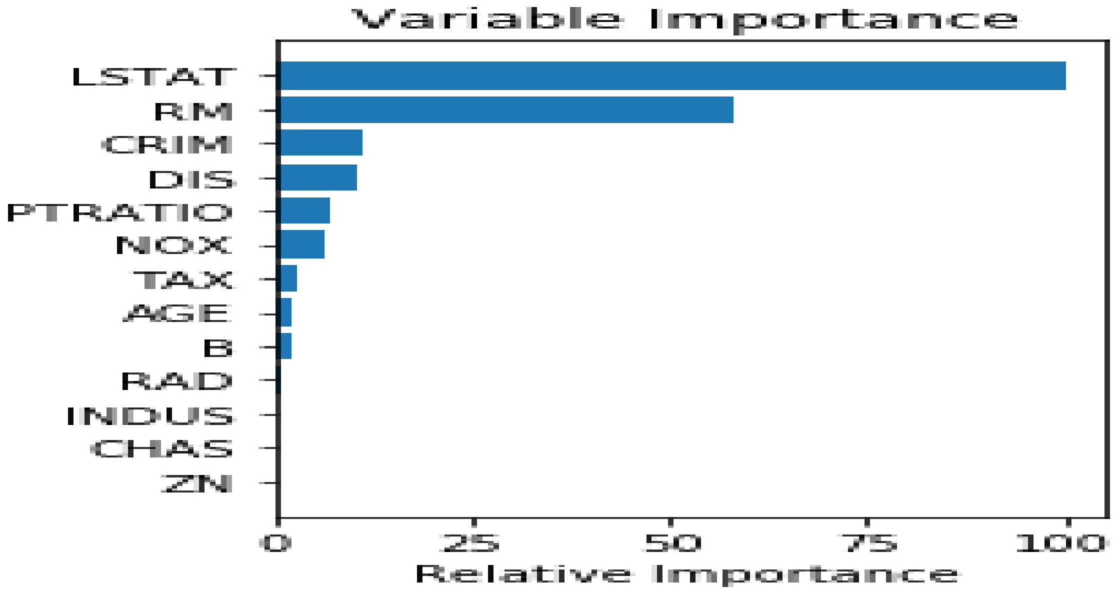
The dataset is then divided into training dataset and test dataset, with 70% : 30% in proportion.

MEDV - Median value of owner-occupied homes in $1000's is defined as the target variable.

Data

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.



3. Methodology

3. 1 Exploratory Data Analysis

For Exploratory Analysis, analysis of data statistics (df.describe()) and correlation between independent variables are conducted.

It can be noted that there are great distance between the minimum and maximum values among some indenpendent variables. For instance, for B, the min value is 0.320000 while the max is 396.899994; for CRIM, the min value is 0.006320 while the max is 88.976196. This can also be demonstrated by the standard deviation (std): 168.537109 for TAX, and 91.294823 for B.

There are also high correlation between certain independent variables, especially between NOX, TAX and INDUS with other variables.

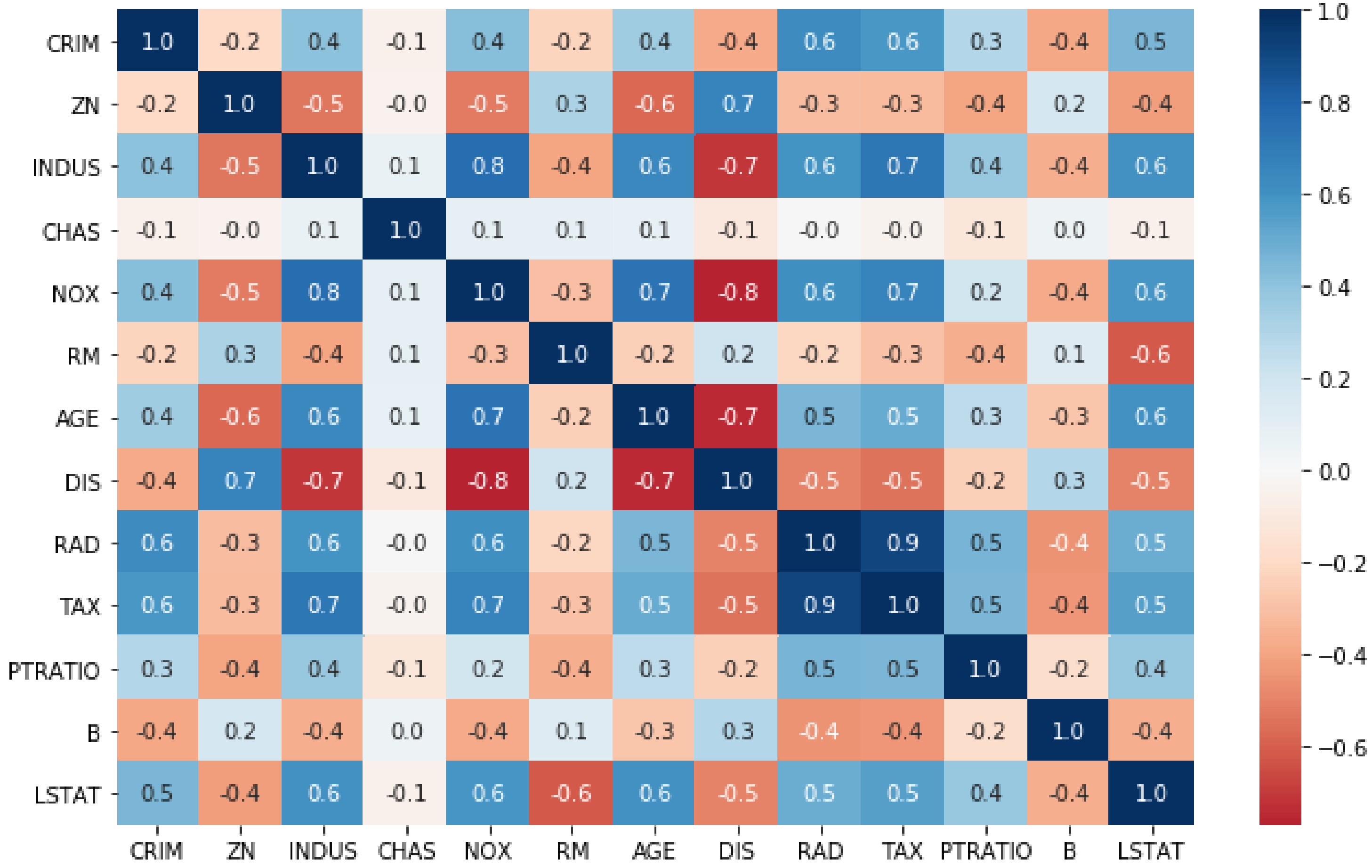
The correlation of NOX with INDUS: 0.8, with AGE: 0.7, and with TAX: 0.7.

The correlation of TAX with INDUS: 0.7, with RAD: 0.9, with CRIM: 0.6, and with NOX: 0.7.

The correlation of INDUS with NOX: 0.8, with LSTAT: 0.6, and with TAX: 0.7.

3. Methodology

3. 1 Exploratory Data Analysis - Correlation between independent variables.

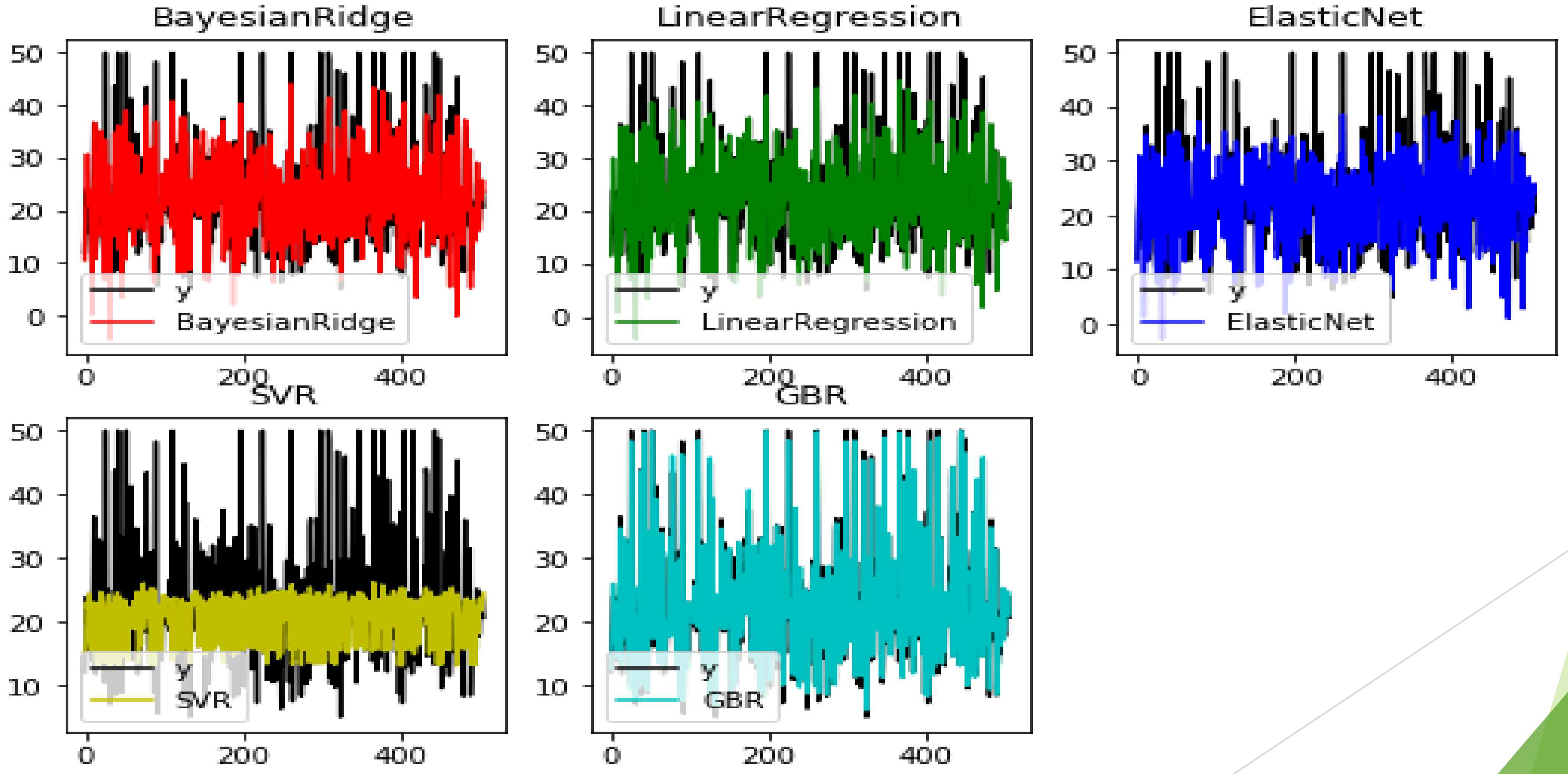


	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
CRIM	1.0	-0.2	0.4	-0.1	0.4	-0.2	0.4	-0.4	0.6	0.6	0.3	-0.4	0.5
ZN	-0.2	1.0	-0.5	-0.0	-0.5	0.3	-0.6	0.7	-0.3	-0.3	-0.4	0.2	-0.4
INDUS	0.4	-0.5	1.0	0.1	0.8	-0.4	0.6	-0.7	0.6	0.7	0.4	-0.4	0.6
CHAS	-0.1	-0.0	0.1	1.0	0.1	0.1	0.1	-0.1	-0.0	-0.0	-0.1	0.0	-0.1
NOX	0.4	-0.5	0.8	0.1	1.0	-0.3	0.7	-0.8	0.6	0.7	0.2	-0.4	0.6
RM	-0.2	0.3	-0.4	0.1	-0.3	1.0	-0.2	0.2	-0.2	-0.3	-0.4	0.1	-0.6
AGE	0.4	-0.6	0.6	0.1	0.7	-0.2	1.0	-0.7	0.5	0.5	0.3	-0.3	0.6
DIS	-0.4	0.7	-0.7	-0.1	-0.8	0.2	-0.7	1.0	-0.5	-0.5	-0.2	0.3	-0.5
RAD	0.6	-0.3	0.6	-0.0	0.6	-0.2	0.5	-0.5	1.0	0.9	0.5	-0.4	0.5
TAX	0.6	-0.3	0.7	-0.0	0.7	-0.3	0.5	-0.5	0.9	1.0	0.5	-0.4	0.5
PTRATIO	0.3	-0.4	0.4	-0.1	0.2	-0.4	0.3	-0.2	0.5	0.5	1.0	-0.2	0.4
B	-0.4	0.2	-0.4	0.0	-0.4	0.1	-0.3	0.3	-0.4	-0.4	-0.2	1.0	-0.4
LSTAT	0.5	-0.4	0.6	-0.1	0.6	-0.6	0.6	-0.5	0.5	0.5	0.4	-0.4	1.0

3. Methodology

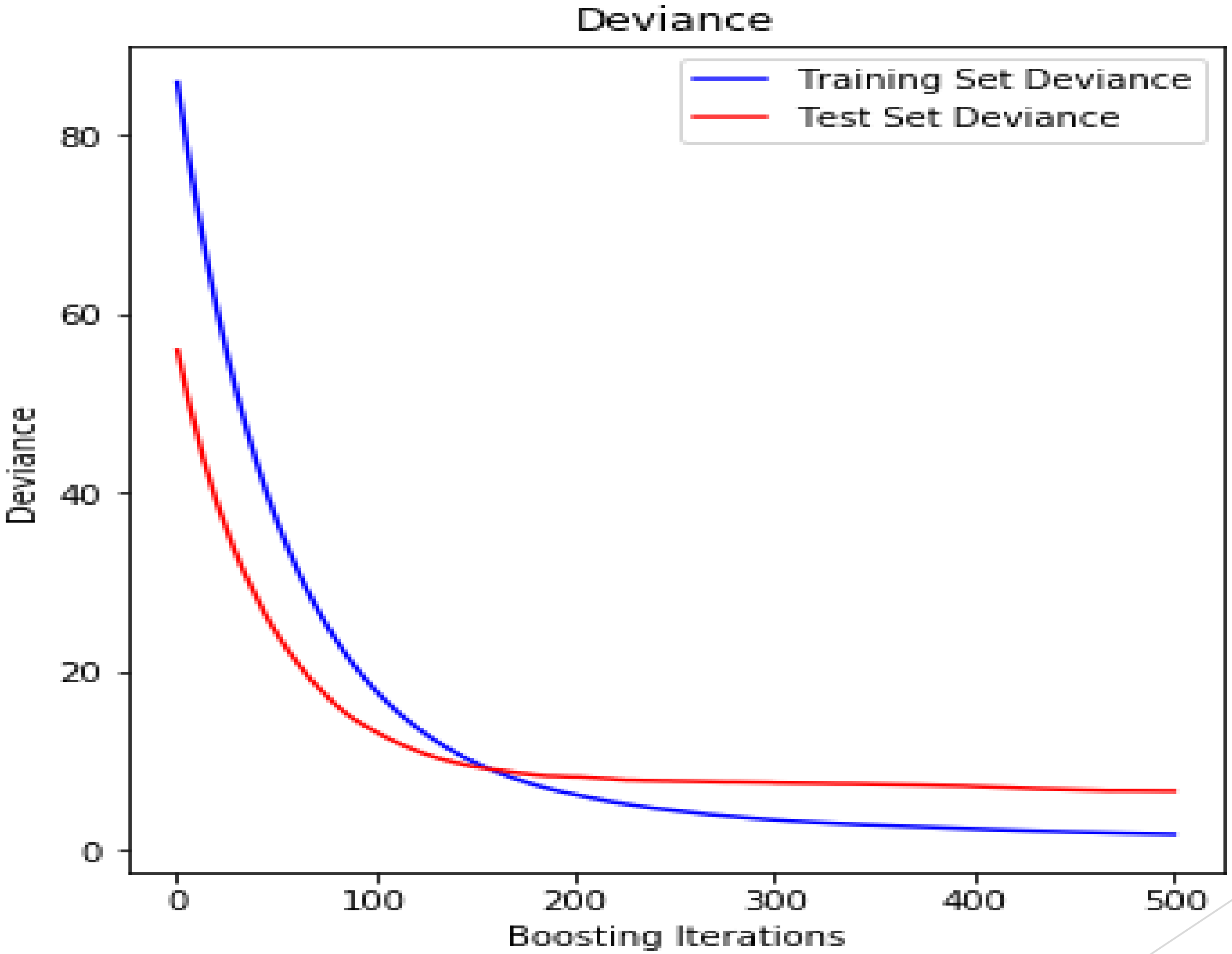
3.2 Predictive Modeling

GradientBoostingRegressor provides a very useful predictive model as it performs much better in predicting the house values than BayesianRidge, LinearRegression, ElasticNet, GradientBoostingRegressor and SVR.



3. Methodology

3.2 Predictive Modeling - Training dataset deviance in comparison test dataset deviance.



Results

Our analysis shows that The above analysis demonstrate that the most important variables for predicting the target variable are LSTAT, RM, CRIM,DIS, PTRATIO, NOX, TAX, and AGE.

The metrics of the predicting models indicate that GBR is best model, as shown by its least mse of 2.014201 and its largest R-square of 0.976141, while SVR is best model as shown by the largest mse of 66.818898 and the smallest R-square of 0.208490 associated with it.

Discussion

The high correlation between certain independent variables can cause problems for predicting models, especially linear regression, which assumes there is no correlation between consecutive residuals. Thus linear regression may not be a optimum model

Further treatment of the independent variables can be conducted to remove the correlations among the independent variables to improve performance of the predicting models

Conclusion

Analysis of the Boston house prices dataset indicates that the most effective factors for predicting the house value are LSTAT - % lower status of the population, RM - average number of rooms per dwelling, CRIM - per capita crime rate by town, DIS - weighted distances to five Boston employment centres, PTRATIO - pupil-teacher ratio by town, NOX - nitric oxides concentration (parts per 10 million), TAX - full-value property-tax rate per $10,000, and AGE - proportion of owner-occupied units built prior to 1940.

GradientBoostingRegressor provides a very useful predictive model as it performs much better in predicting the house values than BayesianRidge, LinearRegression, ElasticNet, and SVR