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.

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.

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

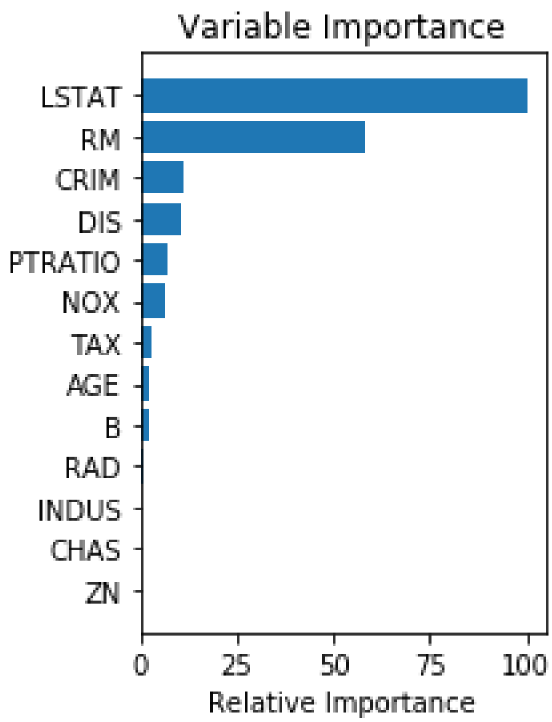
sklearn.datasets of python also contains the dataset, which is used for this project.

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.



- Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

3. Methodology

3. 1 Exploratory Data Analysis

Results - Exploratory 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.

	CRIM	**ZN**	**INDUS**	**CHAS**	**NOX**	**RM**	**AGE**	**DIS**
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136795	0.069170	0.554695	6.284637	68.574898	3.795046
std	8.601545	23.322399	6.860352	0.253993	0.115878	0.702617	28.148859	2.105710
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.074999	5.188425
max	88.976196	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500

Table 1. statistics of variables data

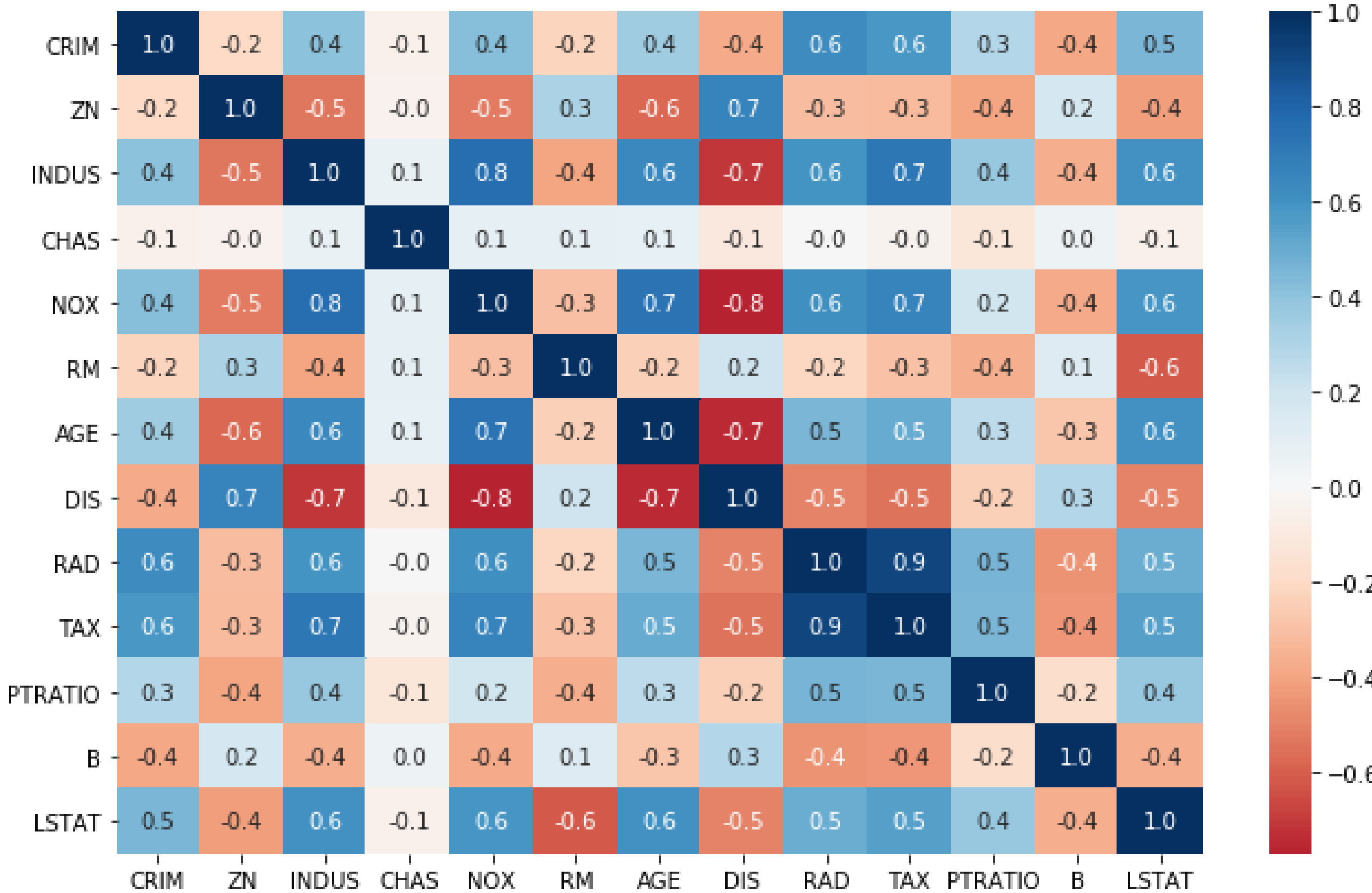


Figure 1. Correlation between independent variables.

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.

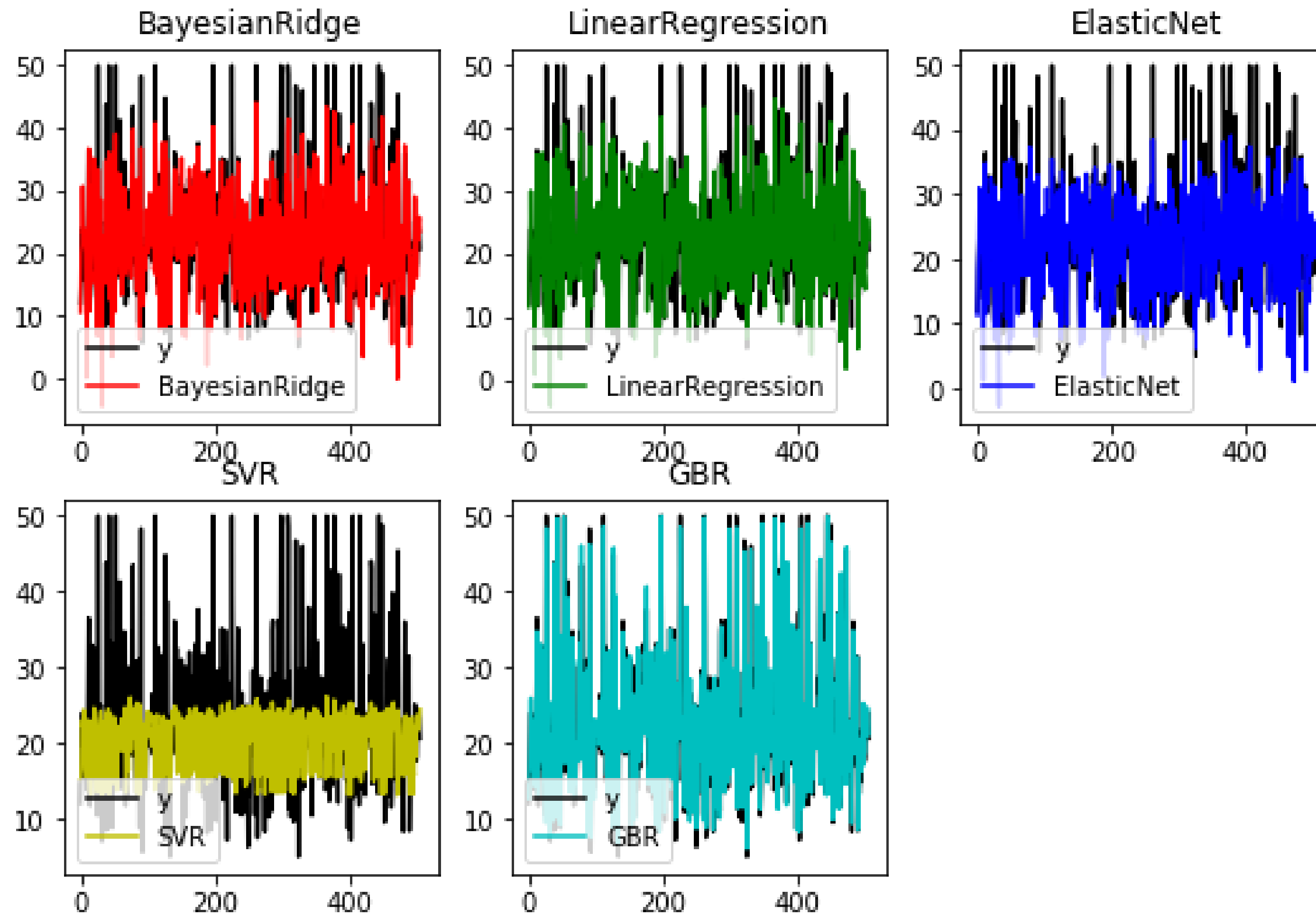


Figure 2. Distribution of actual and predicted improvement using linear regression with equal weights of samples.

Cross-validation score:

	0	1	2	3	4
BayesianRidge	0.706012	0.709296	0.632994	0.778066	0.650570
LinearRegression	0.731768	0.717604	0.630617	0.789995	0.652510
ElasticNet	0.665218	0.689513	0.633569	0.699097	0.638720
SVR	0.249776	0.205957	0.203228	0.233234	0.076399
GBR	0.908896	0.781829	0.757715	0.920407	0.905126

Metrics:

	ev	mae	mse	r2
BayesianRidge	0.731209	3.317905	22.691182	0.731209
LinearRegression	0.740643	3.270866	21.894831	0.740643
ElasticNet	0.686102	3.594529	26.499144	0.686102
SVR	0.260937	5.259882	66.818898	0.208490
GBR	0.976141	1.114241	2.014201	0.976141

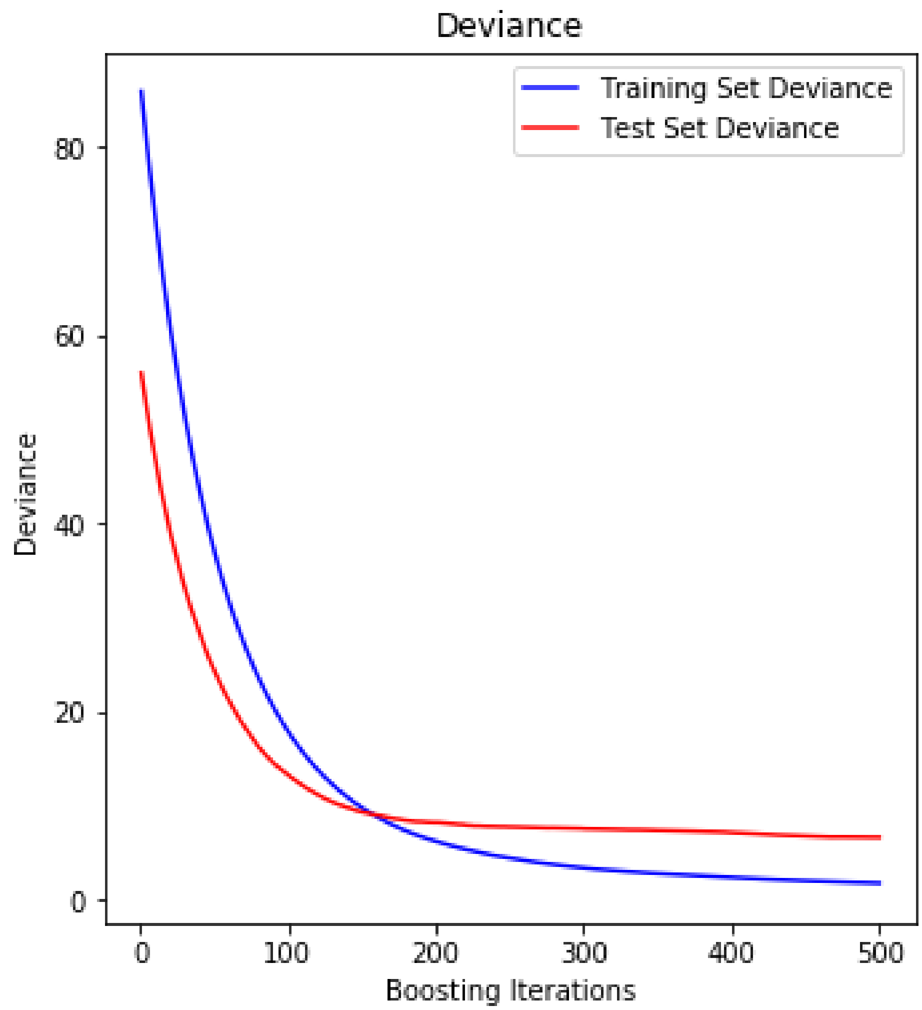


Figure 3. Training dataset deviance in comparison test dataset deviance.

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.