Boston Housing Price Prediction

Pruthvi Ranjan Reddy Pati

SUMMARY

INTRO

The Boston Housing Data is a dataset from 1978 which describes the housing values in Suburbs of Boston using 506 observations of 14 variables. The median value of houses was predicted using 13 independent variables.

APPROACH

To predict the median house values different supervised machine learning models like Linear Regression, Regression Tree, Bagging, Random Forest , Boosting Tree, GAM and Neural Network is used. The data is divided into train and test. Every model is then trained using in sample data and the hyper-parameters are tuned. A final model is selected comparing all in sample MSE of models. The selected final model is then tested to find out of sample MSE.

MAJOR FINDINGS

* Linear Regression performed the worst with an in sample MSE of 24.8
* Boosting tree gave the best in sample MSE of 1.26
* Linear Regression, Regression tree, Bagging, GAM, Neural Network Random Forest, Boosting tree is the order of in sample performance from worst to best.
* The out of sample MSE of boosting is 4.6 significantly higher compared to the in sample MSE. It can be a sign of overfitting, but is still better when compared to other models out of sample error

**Exploratory Data Analysis**

The *Boston Housing Data* is a dataset from 1978 which was used to predict the median value of houses based on 13 independent variables. In this section each variable is explored and described using its mean, median, and standard deviation values. The pairwise correlations of each variable are also explored and featured in a graph. Below are tables that feature the summary statistics for each variable. Prior to any analysis being completed the data was split into a training subset and a testing subset to be used later. 75% of the data points were put into the training subset.

***Variables (Table 1.1)***

|  |  |
| --- | --- |
| ***Variable*** | ***Description*** |
| ***crim*** | Per capita crime rate by town |
| ***zn*** | Proportion of residential land for lots >25,000sq ft |
| ***indus*** | Proportion of non-retail biz acres per town |
| ***chas*** | Dummy variable for if tract bounds river or not |
| ***nox*** | Nitrogen Oxides ppm measurement |
| ***rm*** | Avg. number of rooms per dwelling |
| ***age*** | Proportion of owner-occupied units built before 1940 |
| ***dis*** | Weighted mean of distances to 5 Boston employment centers |
| ***rad*** | Index of access to radial highways |
| ***tax*** | Full value property tax rate per $10,000 |
| ***ptratio*** | Pupil-Teacher ratio by town |
| ***black*** | 1000\*(Bk – 0.63)2 where Bk is the proportion of blacks per town |
| ***lstat*** | Lower status of the population (percent) |
| ***medv*** | Median value of owner-occupied houses in $1,000s. The dependent variable in this study. |

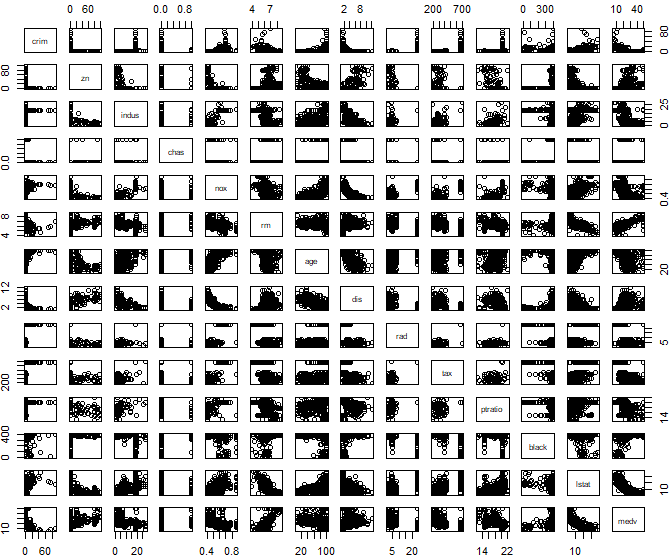
***Summary Statistics (Table 1.2)***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***crim*** | ***zn*** | ***indus*** | ***chas*** | ***nox*** | ***rm*** | ***age*** |
| **Mean** | 3.63 | 11.04 | 10.95 | 0.07 | 0.55 | 6.27 | 68.95 |
| **Median** | 0.21 | 0 | 9.69 | 0 | 0.53 | 6.18 | 77.70 |
| **Std. Dev.** | 9.04 | 23.23 | 6.88 | 0.26 | 0.11 | 0.69 | 27.78 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***dis*** | ***rad*** | ***tax*** | ***ptratio*** | ***black*** | ***lstat*** | ***medv*** |
| **Mean** | 3.83 | 9.24 | 402.90 | 18.45 | 359.76 | 12.67 | 22.59 |
| **Median** | 3.27 | 5.00 | 329.00 | 19.00 | 392.05 | 11.28 | 21.20 |
| **Std. Dev.** | 2.12 | 8.61 | 168.13 | 2.13 | 87.96 | 7.14 | 9.32 |

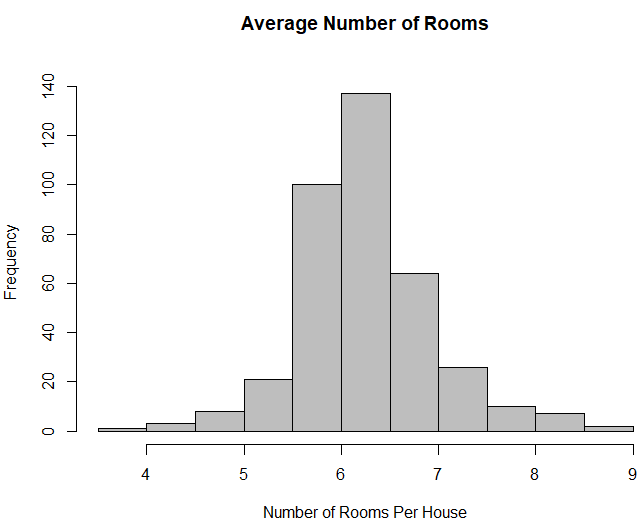
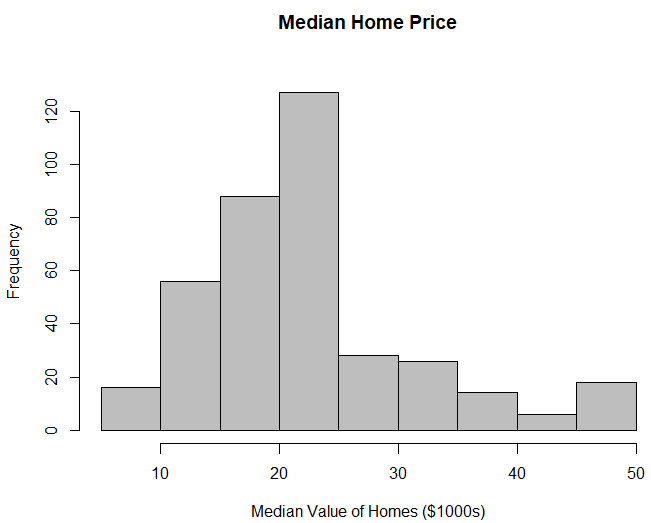
From Table 1.2 it is evident that some of the variables such as crime rate and the index of access to radial highways have a large degree of variability. In general there is a positive skew within the data, but some variables such as age and black show a negative skew. In the next part of the exploratory data analysis the pairwise correlation is examined in the form of a pairwise correlation plot.

***Fig. 1.1***

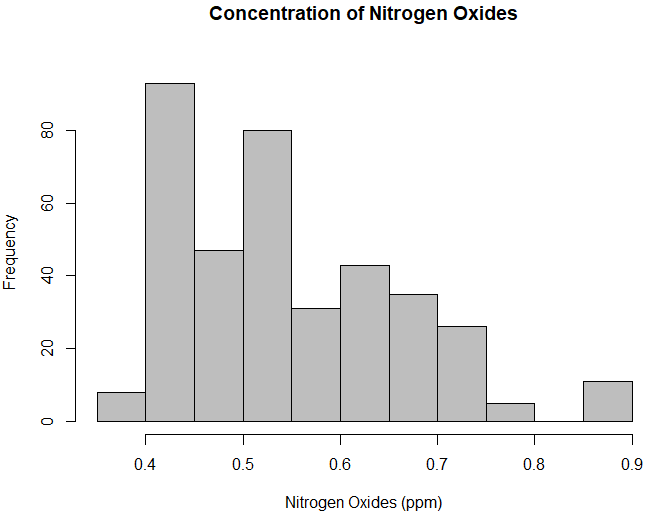


The pairwise correlation plot is only one step in visualizing the data. Data visualization is an important step into understanding the dataset because it gives the user intuition regarding the relationship of variables to each other. In addition to pairwise plots, histograms are another good way to show data distributions. Since there are 14 variables and too many to individually plot a few have been selected. These are pictured below. In Fig. 1.2 and 1.3 it is shown that the median home price and avg. number of rooms roughly follow a normal distribution while dist and nox variables follow an exponential and bimodal distribution, respectively.

***Fig 1.2 Fig. 1.3***



***Fig. 1.4 Fig. 1.5***



### 1.Linear Regression

#### Finding the best model:

To choose the best model, we performed variable selection using three techniques:

* Best AIC
* Best BIC
* Lasso Regression

**AIC**

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 39.902927454 5.991578091 6.659836 1.000086e-10  
## lstat -0.564911286 0.056283669 -10.036860 4.103512e-21  
## rm 3.613851697 0.472637655 7.646136 1.823375e-13  
## ptratio -0.985298585 0.156201618 -6.307864 8.121580e-10  
## dis -1.586169887 0.226436333 -7.004927 1.182407e-11  
## nox -17.461328914 4.178615136 -4.178736 3.665602e-05  
## black 0.008513113 0.003187276 2.670968 7.899040e-03  
## rad 0.333226900 0.075051787 4.439960 1.191150e-05  
## crim -0.128131790 0.040376876 -3.173395 1.633457e-03  
## tax -0.012641181 0.004002691 -3.158170 1.718811e-03  
## zn 0.038737674 0.016669449 2.323873 2.067631e-02

**BIC**

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 39.2513063 5.44151402 7.213306 3.069268e-12  
## lstat -0.6186808 0.05599277 -11.049298 9.592426e-25  
## rm 3.9432681 0.47637134 8.277719 2.262576e-15  
## ptratio -1.0098269 0.13391665 -7.540712 3.593515e-13  
## dis -1.2995690 0.20565373 -6.319210 7.501054e-10  
## nox -18.0764204 3.81740912 -4.735259 3.113136e-06

**Lasso Regression**

## 14 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 18.194198689  
## crim -0.021202247  
## zn .   
## indus .   
## chas 0.549460686  
## nox -1.130957109  
## rm 3.995298752  
## age .   
## dis -0.265196592  
## rad .   
## tax .   
## ptratio -0.739804443  
## black 0.004721052  
## lstat -0.557233228

* Below is a summary of the MSE of the three models:

## MSE  
## AIC model 24.81854  
## BIC model 26.69307  
## Lasso model 27.83522

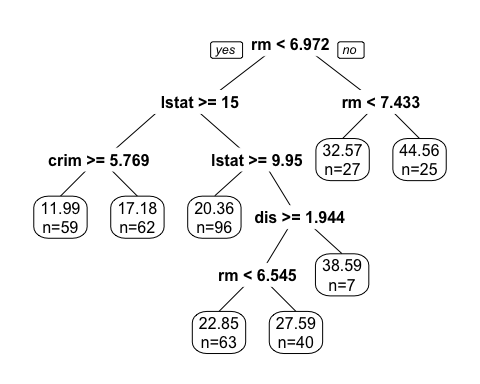
* We chose the step AIC model as the best model, because it provides the least MSE out of the three models.
* In Sample metrics:

## Final model  
## MSE 24.8185378  
## R-Squared 0.7173463  
## Adjusted R-Squared 0.7096655

### 2.Regression tree

* We fit a Regression Tree on our training data. Below is the output of the model:

## n= 379   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 379 32312.4000 22.51821   
## 2) rm< 6.9715 327 13533.2000 20.00306   
## 4) lstat>=15 121 2315.8210 14.65289   
## 8) crim>=5.76921 59 637.1485 11.99492 \*  
## 9) crim< 5.76921 62 865.1905 17.18226 \*  
## 5) lstat< 15 206 5719.4310 23.14563   
## 10) lstat>=9.95 96 563.5074 20.36146 \*  
## 11) lstat< 9.95 110 3762.3240 25.57545   
## 22) dis>=1.944 103 1637.5020 24.69126   
## 44) rm< 6.5445 63 629.7575 22.85079 \*  
## 45) rm>=6.5445 40 458.2360 27.59000 \*  
## 23) dis< 1.944 7 859.4286 38.58571 \*  
## 3) rm>=6.9715 52 3702.3580 38.33462   
## 6) rm< 7.433 27 809.1919 32.57407 \*  
## 7) rm>=7.433 25 1029.5620 44.55600 \*

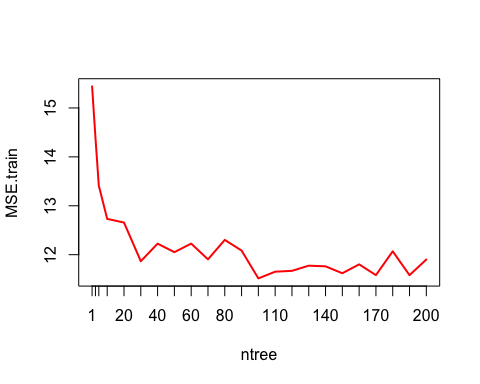


## [1] 15.44069

* The In-sample MSE of the Regression Tree is 15.44069.

### 

### 3.Bagging forest

* Finding the optimal number of trees 
* Out of bag sample(OOB) error

## [1] 18.4976

* In Sample MSE:

## [1] 12.0329

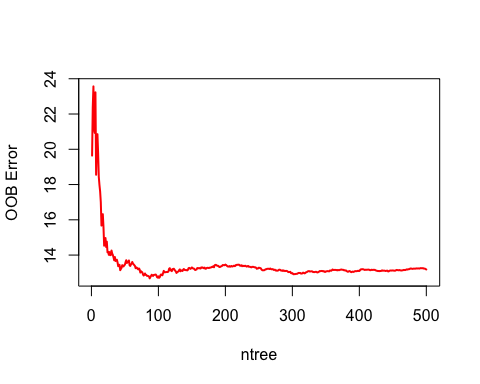
* The In-sample MSE of the Bagging Forest is 12.0329.

### 4.Random forest

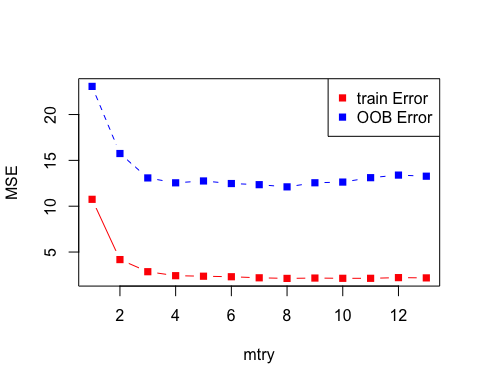
##   
## Call:  
## randomForest(formula = medv ~ ., data = boston\_train, importance = TRUE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 4  
##   
## Mean of squared residuals: 13.18279  
## % Var explained: 84.54

* Variable importance

## %IncMSE IncNodePurity  
## crim 7.6988676 1911.5471  
## zn 0.7036219 172.3154  
## indus 5.8034303 1938.5613  
## chas 0.2367873 134.4321  
## nox 8.7921401 1767.9402  
## rm 34.2420691 9548.7940  
## age 3.4372218 1026.2992  
## dis 7.0059858 2240.4870  
## rad 1.0275634 253.8919  
## tax 3.0139253 959.7273  
## ptratio 7.2125874 1946.8615  
## black 1.5761417 710.0267  
## lstat 55.0168040 9049.4047



## 1 2 3 4 5 6 7 8 9 10 11 12 13

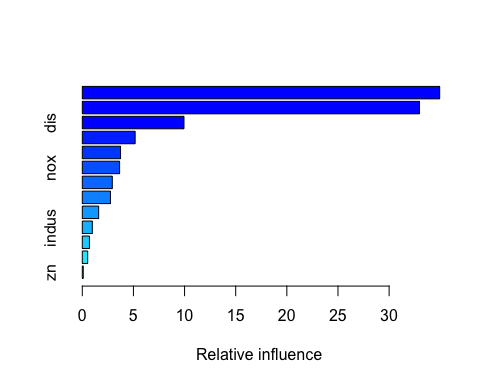


* In-sample mse:

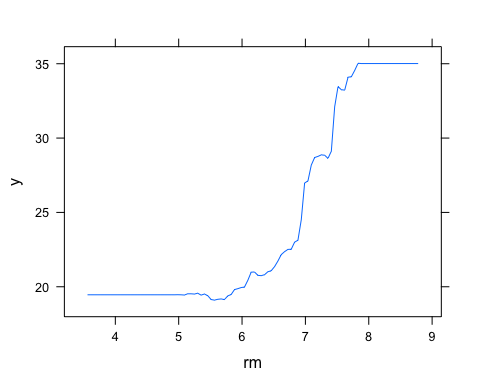
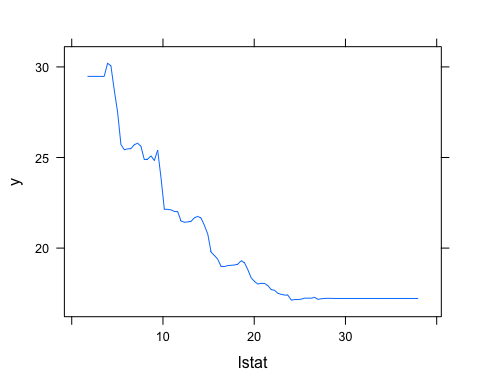
## [1] 2.346243

* The In-sample MSE of the Random Forest is 2.0329.

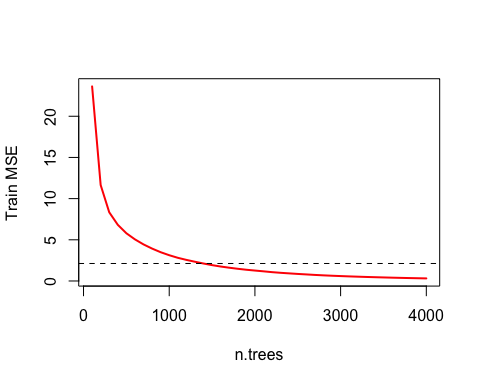
### 5.Boosting forest



## var rel.inf  
## lstat lstat 34.94468873  
## rm rm 32.96859136  
## dis dis 9.94339352  
## crim crim 5.16721336  
## age age 3.74137787  
## nox nox 3.64735877  
## black black 2.93251509  
## ptratio ptratio 2.75455609  
## tax tax 1.60039438  
## indus indus 0.97921467  
## rad rad 0.69982285  
## chas chas 0.52836572  
## zn zn 0.09250758



### In sample prediction



## [1] 1.266747

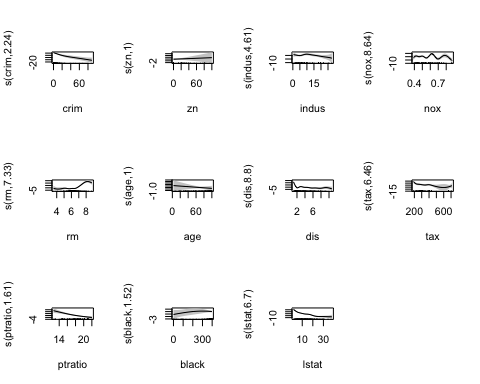
* The In-sample MSE of the Boosting Forest is 1.324547.

### 

### 6.GAM

* Finding the optimal number of trees

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## medv ~ s(crim) + s(zn) + s(indus) + chas + s(nox) + s(rm) + s(age) +   
## s(dis) + rad + s(tax) + s(ptratio) + s(black) + s(lstat)  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 18.3110 1.5536 11.786 < 2e-16 \*\*\*  
## chas 0.5785 0.7170 0.807 0.42040   
## rad 0.4338 0.1612 2.692 0.00747 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df F p-value   
## s(crim) 2.242 2.800 18.150 2.43e-09 \*\*\*  
## s(zn) 1.000 1.000 0.275 0.600   
## s(indus) 4.608 5.534 1.305 0.153   
## s(nox) 8.636 8.935 11.576 3.71e-16 \*\*\*  
## s(rm) 7.335 8.305 19.617 < 2e-16 \*\*\*  
## s(age) 1.000 1.000 0.542 0.462   
## s(dis) 8.805 8.982 7.108 1.70e-09 \*\*\*  
## s(tax) 6.462 7.495 5.267 5.64e-06 \*\*\*  
## s(ptratio) 1.613 2.003 13.254 2.84e-06 \*\*\*  
## s(black) 1.517 1.862 1.309 0.357   
## s(lstat) 6.702 7.831 19.121 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.874 Deviance explained = 89.2%  
## GCV = 12.496 Scale est. = 10.751 n = 379



* In-sample MSE of GAM

## [1] 9.250207

* The In-sample MSE of the GAM is 9.25.

### 7.Neural Networks

* In-Sample MSE of neural net

## [1] 5.21627

* The In-sample MSE of the Neural Network is 4.6632. ###In sample model comparisions

## Model MSE  
## 1 Linear Regression 24.818538  
## 2 Regression Tree 15.440691  
## 3 Bagging Forest 18.497597  
## 4 Random forest 2.346243  
## 5 Boosting 1.266747  
## 6 GAM 9.250207  
## 7 Neural Network 5.216270

* Final model we have chosen is Boosting based on MSE

## Data MSE  
## 1 In\_Sample 1.266747  
## 2 Out\_of\_Sample 4.594178

* The out of sample error is comparitively greater than in-sample mse, but is better than other models.