Project 2:
Examining Housing Prices in Boston

STAT 6620: Statistical Learning with R

By

Rakesh Sunkari

Net ID: re7332

Boston Housing Dataset

The dataset used in this project comes from a paper written on the relationship between house prices and clean air in the late 1970's by David Harrison of Harvard and Daniel Rubinfeld of University of Michigan.The dataset is downloaded from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Housing) and concerns with housing values in suburbs of Boston.

Reading the data

This dataset has 506 observations of 14 attributes. The following table gives information about the attributes in the dataset.

Variable Name	***Definition***
CRIM	Per capita crime rate
ZN	Perc of land zoned for lots
INDUS	Proportion of non-retail 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 centers
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per USD 10,000
PTRATIO	pupil-teacher ratio by town
MEDV	median value of owner-occupied homes in USD 1000's
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population

From this dataset, I would like to explore the housing dataset with aid of R Statistical package. I would like to see which attributes affect the housing prices across Boston using some of the well-known machine learning algorithms.

Exploring the data

The given dataset is based on the house prices. So, I'm interested to apply summary statistics to two variables MEDV and NOX

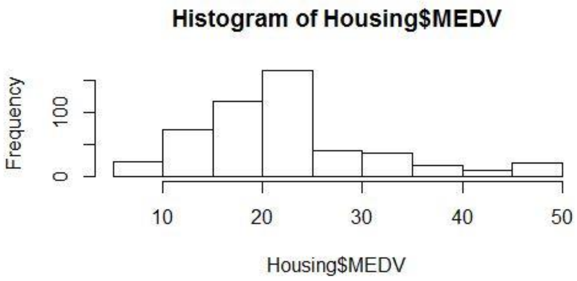
- Distribution of Housing prices in USD 1000.

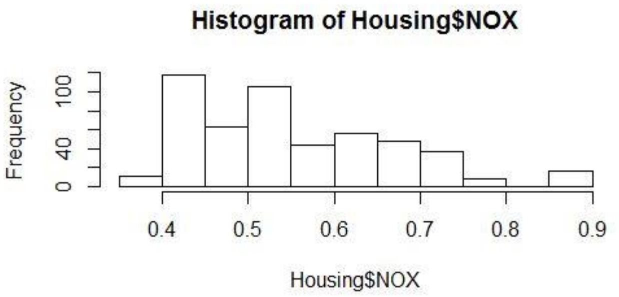
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
5.00   17.02   21.20   22.53   25.00   50.00
```

- Distributon of Nitrous concentration in Boston area.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.3850  0.4490  0.5380  0.5547  0.6240  0.8710
```

We are going to apply graphical representation for the given variables.





From the above histogram, we can see that the MEDV and NOX are not normally distributed, they are right skewed.

Training and testing the data using linear regression

We are going to apply linear regression algorithm to the data as MEDV is predicted variable using RM,LSTAT,CRIM,ZINC,,CHAS,DIS as explanatory variables.

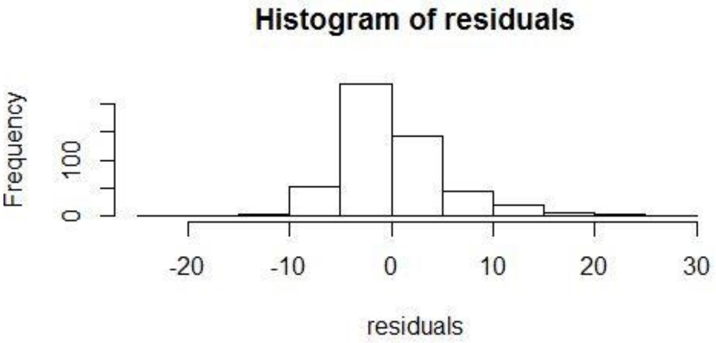
The R output is as follows:

```
Call:
lm(formula = MEDV ~ RM + LSTAT + CRIM + ZN + CHAS + DIS, data = Housing)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.8421  -2.9696  -0.8929   1.8371  25.9671 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.01209    3.29811   1.823 0.068918 .  
RM           4.45430    0.43297  10.288  < 2e-16 ***
LSTAT       -0.65100    0.04927 -13.213  < 2e-16 ***
CRIM        -0.13813    0.03126  -4.419 1.22e-05 ***
ZN           0.06868    0.01381   4.973 9.09e-07 ***
CHAS         3.42915    0.92832   3.694 0.000245 ***
DIS         -0.98925    0.16593  -5.962 4.73e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.214 on 499 degrees of freedom
Multiple R-squared:  0.6824,Adjusted R-squared:  0.6786 
F-statistic: 178.7 on 6 and 499 DF,  p-value: < 2.2e-16
```

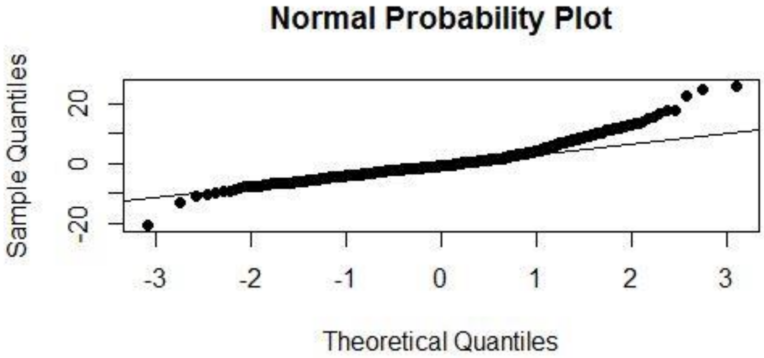
From the output we can see thatCrime rate (CRIM), Average number of rooms (RM), distance to 5 nearest employment centers (DIS), affects the housing prices which really seems to be intuitive. But Adjusted R-squared is 0.6786 which is not so good. But this might be due to the effect of other insignificant variables.

The mean of housing data residuals is said to be `[1] 1.17291e-16`

The histogram of residuals



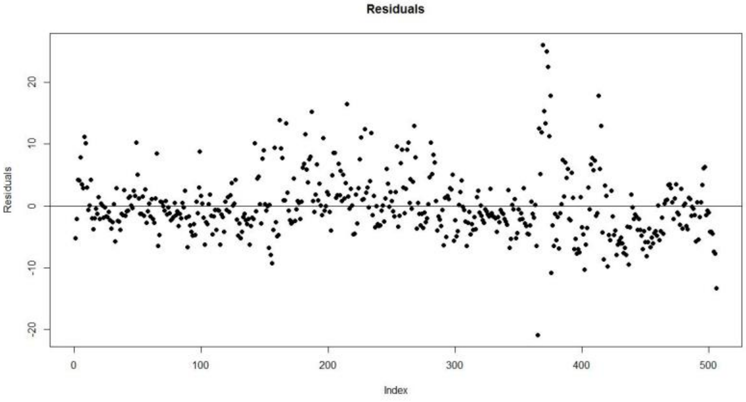
From the histogram we can assume that the residuals are normal and I'm very likely to test the normality of the residuals.



From the above linear regression table, we can see that p-value: < 2.2e-16, therefore we can say that the plot is not normal.

Evaluating the dataset

We are going to check the correlation whether it is necessary to apply transformation

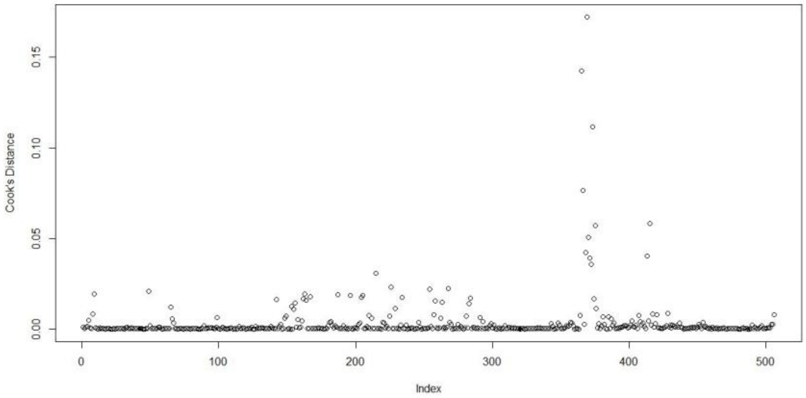


From the Scatterplot, we can come to know that the data need to be transformed.

Checking the cook's distance because Cook's D measures the influence of the *i*th observation on all n fitted values.

The magnitude of D_i is usually assessed as:
 if the percentile value is less than 10 or 20 % than the *i*th observation has little apparent influence on the fitted values

if the percentile value is greater than 50%, we conclude that the *i*th observation has significant effect on the fitted values



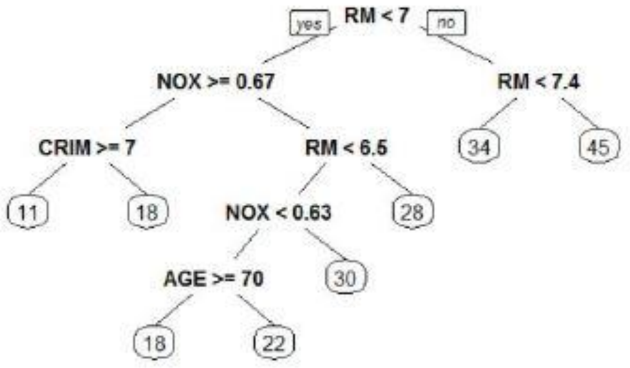
Improving the model performance

We first create two dataset like housing1 and housing2 using linear regression and compare them with anova,

```
Analysis of Variance Table

Model 1: MEDV ~ RM + LSTAT + CRIM + ZN + CHAS + DIS
Model 2: MEDV ~ RM + LSTAT + CRIM + CHAS + DIS
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)    
1    499 13566                                  
2    500 14238 -1   -672.31 24.729 9.089e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the anova table , we can see that the p-value is greater that F value and data is said to be significant.

Using decision tree



From decision tree leaves, we can see that housing prices are affected by Crime rate (CRIM), Pollution (NOX), Average number of rooms (RM) and proportion of owner-occupied units built prior to 1940(AGE).

Conclusion

As we can from our initial analysis that housing prices are dependent on the location. In that case, CART has performed better compared to linear regression. But when we used all the variables, we can see from the above table that linear regression outperforms Decision Trees. Cross- Validation has improved the basic CART to a great extent, but it still underperforms compared to linear regression.

So, factors appears to be affecting the housing prices in Boston area according to this analysis are
Per capita crime-rate of that town (CRIM)
Location of that area
Average number of rooms in that dwelling (RM)
Pollution i.e. NOX concentration (NOX)
Distance of that area from 5 nearest employment centers (DIS)

Appendix : R code

```
#Boston dataset from UCI repository

#step1 : Reading the data

Housing <- read.csv("Housing.csv", stringsAsFactors = TRUE)

str(Housing)

names(Housing)

#step2 : exploring the data

summary(Housing)

summary(Housing$MEDV)

summary(Housing$NOX)

hist(Housing$MEDV)

hist(Housing$NOX)

#step3 : creating training and test datasets

#linear regression

Housing<-lm(MEDV ~ RM + LSTAT + CRIM + ZN + CHAS + DIS, data= Housing)

summary(Housing)

#checking for normality

mean(Housing$residuals)

hist(Housing$residuals, xlab="residuals", main="Histogram of residuals")
```

```
qqnorm(Housing$residuals, main="Normal Probability Plot", pch=19)

qqline(Housing$residuals)

#step4 : Evaluating the data

#checking for the correlation

plot(Housing$residuals, main="Residuals", ylab="Residuals", pch=19)

abline(h=0)

#checking influential observations by using cooksd method

cd=cooks.distance(Housing)

plot(cd, ylab="Cook's Distance")

abline(h=qf(c(.2,.5),6, 499))

ic=(1:506)[cd>qf(c(.2,.5), 6,499)]

text(ic,cd[ic], as.character(Housing$OBS [ic]),adj=c(1,1))

#step5 : Improving the data

#Using ANOVA table

Housing1<-lm(MEDV ~ RM + LSTAT + CRIM + ZN + CHAS + DIS)

Housing2<-lm(MEDV ~ RM + LSTAT + CRIM + CHAS + DIS)

anova(bost1,bost2)
```