Predicting Boston Housing Prices?

Nathan Lee and Tejas Patel

2023-05-24

Project Overview

The main purpose of this report is to investigate whether it is possible to predict the median value of owner-occupied homes in Boston based on given features. The guiding question for this analysis is: What are the effects of variables such as crime rate, proportion of residential land zoned, average number of rooms, accessibility to highways, etc. on the median value of owner-occupied homes in Boston?

The analysis will be conducted using a dataset from Kaggle that contains information on various features of houses in Boston. The dataset includes variables such as per capita crime rate by town, proportion of residential land zoned for lots over 25,000 sq.ft., average number of rooms per dwelling, index of accessibility to radial highways, and more. The hypothesis is that a higher crime rate (CRIM), a higher nitric oxide concentration (NOX), and more rooms (RM) will affect MEDV the most.

The purpose of this analysis is to develop a regression model that accurately predicts the median value of owner-occupied homes in Boston based on these features. The report will include exploratory data analysis, data preparation, model development and evaluation, prediction and conclusion.

Explaining the Data

The data for this analysis was obtained from Kaggle. The dataset contains information on various features of houses in Boston. The variables used in this analysis are as follows:

1) CRIM: This variable measures the per capita crime rate by town.

(i) Data Type: numeric
(ii) Range: 0.00632 - 88.9762

2) ZN: This variable measures the proportion of residential land zoned for lots over 25,000 square feet.

(i) Data Type: numeric
(ii) Range: 0 - 100

3) INDUS: This variable measures the proportion of non-retail business acres per town.

(i) Data Type: numeric
(ii) Range: 0.46 - 27.74

4) CHAS: This variable is a Charles River dummy variable (1 if tract bounds river, 0 otherwise).

(i) Data Type: categorical

(ii) Levels: 0, 1

5) NOX: This variable measures the nitric oxide concentration (parts per 10 million).

(i) Data Type: numeric
(ii) Range: 0.385 - 0.871

6) RM: This variable measures the average number of rooms per dwelling.

(i) Data Type: numeric
(ii) Range: 3.561 - 8.780

7) AGE: This variable measures the proportion of owner-occupied units built prior to 1940.

(i) Data Type: numeric
(ii) Range: 2.9 - 100.0

8) DIS: This variable measures the weighted distances to five Boston employment centres.

(i) Data Type: numeric
(ii) Range: 1.1296 - 12.1265

9) RAD: This variable measures the index of accessibility to radial highways.

(i) Data Type: numeric
(ii) Range: 1 - 24

10) TAX: This variable measures the full-value property-tax per $10,000.

(i) Data Type: numeric
(ii) Range: 187 - 711

11) PTRATIO: This variable measures the pupil-teacher ratio by town.

(i) Data Type: numeric
(ii) Range: 12.6 - 22.0

12) B: This variable is the result of the equation B=1000(Bk - 0.63)^2 where Bk is the proportion of black people by town.

(i) Data Type: numeric
(ii) Range: 0.32 - 396.90

13) LSTAT: This variable measures the % lower status of the population.

(i) Data Type: numeric
(ii) Range: 1.73 - 37.97

Output variable: 1) MEDV (predicted variable): This variable measures the median value of owner-occupied homes in $1000's. (i) Data Type: numeric (ii) Range: 5 - 50

Analysis

Multiple Linear Resgression Our first thought was to use linear regression to assess and quantify the relationship between the dependent variable, MEDV, and independent variables.

Reading in and fitting first Linear regression model using MEDV(Median value of owner-occupied homes in $1000's) as the predicted variable.

```
##       CRIM                ZN             INDUS       CHAS         NOX        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   0:471   Min.   :0.3850  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1: 35   1st Qu.:0.4490  
##  Median : 0.25651   Median :  0.00   Median : 9.69           Median :0.5380  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14           Mean   :0.5547  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10           3rd Qu.:0.6240  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74           Max.   :0.8710  
##        RM             AGE              DIS              RAD        
##  Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000  
##  1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000  
##  Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000  
##  Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549  
##  3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000  
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000  
##       TAX           PTRATIO            B              LSTAT      
##  Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73  
##  1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95  
##  Median :330.0   Median :19.05   Median :391.44   Median :11.36  
##  Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65  
##  3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95  
##  Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97  
##       MEDV      
##  Min.   : 5.00  
##  1st Qu.:17.02  
##  Median :21.20  
##  Mean   :22.53  
##  3rd Qu.:25.00  
##  Max.   :50.00  
```

CRIM, NOX, DIS, TAX, PTRATIO, and LSTAT are the variables that have a negative effect on MEDV, while ZN, INDUS, CHAS, RM, AGE, RAD and B have a positive effect. This makes sense since things like higher crime rate tends to bring property value down while having more rooms in a property tends to bring property value up.

Fitting a Linear Regression model using only those with significant p-values so that we can focus on the variables with the most impact.

```
## 
## Call:
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + 
##     TAX + PTRATIO + B + LSTAT, data = boston.data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -15.5984  -2.7386  -0.5046   1.7273  26.2373 
## 
```

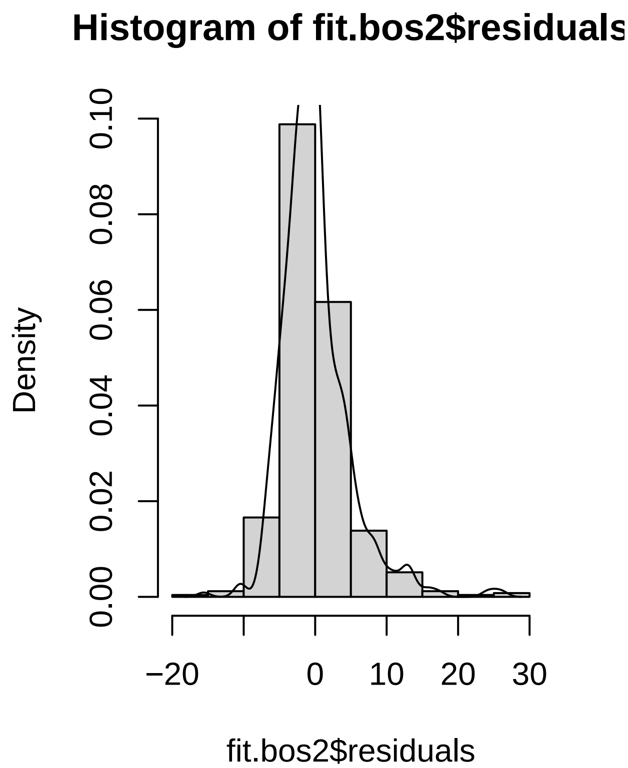
```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## CRIM         -0.108413   0.032779  -3.307 0.001010 **
## ZN            0.045845   0.013523   3.390 0.000754 ***
## CHAS1         2.718716   0.854240   3.183 0.001551 **
## NOX         -17.376023   3.535243  -4.915 1.21e-06 ***
## RM            3.801579   0.406316   9.356  < 2e-16 ***
## DIS          -1.492711   0.185731  -8.037 6.84e-15 ***
## RAD           0.299608   0.063402   4.726 3.00e-06 ***
## TAX          -0.011778   0.003372  -3.493 0.000521 ***
## PTRATIO      -0.946525   0.129066  -7.334 9.24e-13 ***
## B             0.009291   0.002674   3.475 0.000557 ***
## LSTAT        -0.522553   0.047424 -11.019  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

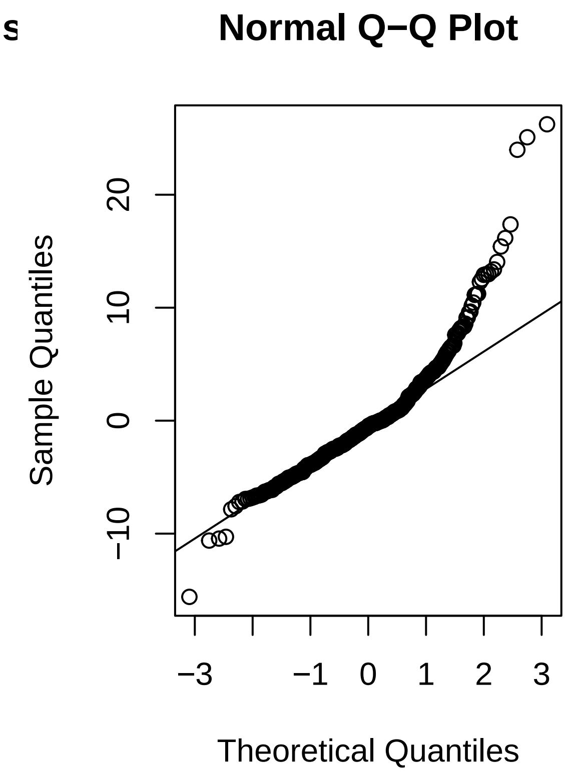
Normality Do a Shapiro test to check normality of residuals.

```
##
##  Shapiro-Wilk normality test
##
## data:  as.numeric(fit.bos2$residual)
## W = 0.90131, p-value < 2.2e-16
```

Since p-value is $< .05$ (p-value: $< 2.2e-16$, therefore $< .05$), the null hypothesis that the data is normally distributed is rejected, and we show that the data is not normally distributed.

Residuals Display a plot of the residuals from the fit.bos2 model that includes a histogram with a density curve and a QQ Plot in the same plot window.





These plots further show that the residuals are not normally distributed, since the histogram is a symmetric bell shaped curve centered at 0, and the Q-Q plot deviates from the straight line.

Predicting with Linear Regression Models We separated the independent variables into categories. The town category was any variable that was by town, the location variable was the distance from the employment centers and weighted distance from highways, and the prop variable is short for property, which were things that had to do directly with the property, so the room number and property tax.

Then calculate and print the RMSE for the models. Note: use the parameter na.rm = TRUE in the call to mean if you have NA values. This tells the mean function to ignore NAs. Otherwise you will not get a real number.

```
## [1] "5.06837  is the predicted town value RMSE."
```

```
## [1] "6.27515  is the predicted location value RMSE."
```

```
## [1] "6.9747  is the predicted property RMSE."
```

From the RMSE outputs, we found that the town that the property was in was the best predictor since it had the lowest RMSE value.

We also wanted to see the impact of the variables that we initially thought would have the most impact on MEDV when we first started looking at the data, which were the CRIM(crime rate), NOX(nitric oxide concentration), and RM(room number) variables.

```
## [1] "7.18504  is the predicted CRIM value RMSE."
```

```
## [1] "6.80628  is the predicted NOX value RMSE."
```

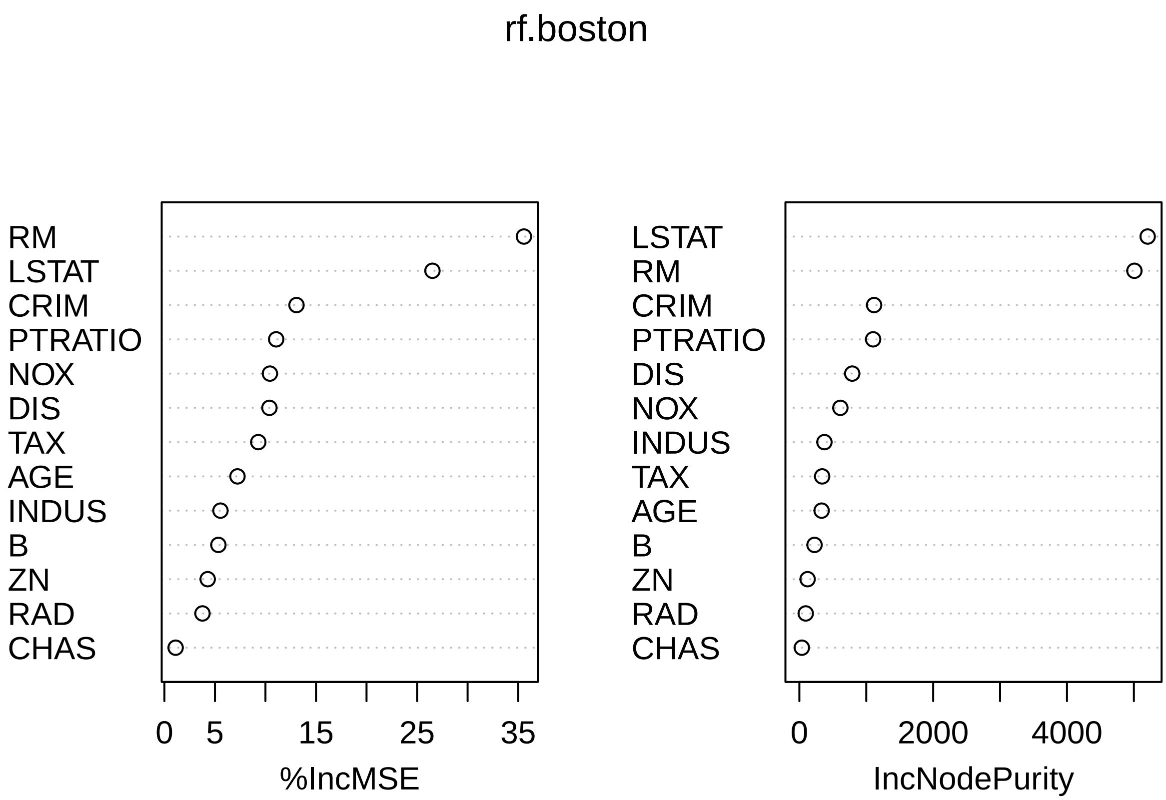
```
## [1] "8.59334  is the predicted RM RMSE."
```

From the output values, we see that the NOX variable had the lowest RMSE value, which means that this variable was the best predictor(between the specific 3 variables here) of MEDV.

Using a Decision Tree Model We also thought to use a decision tree model to see if we can predict the median cost of housing in Boston. This idea was taken from a lab that we did, but we do not go as in-depth, as we just wanted to compare the results from the decision tree and the linear regressions that we did.

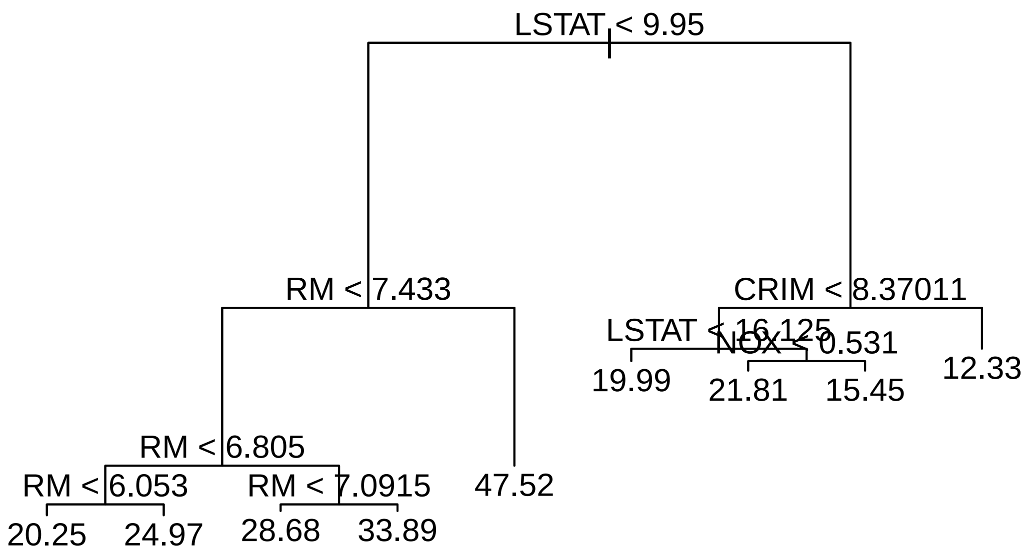
These statements create testing and training sets using half for train and half for test. Also creates the vector of true outcomes, or labels for use in model prediction later.

```
##           %IncMSE IncNodePurity
## CRIM    13.071899    1117.47217
## ZN       4.282928     123.15957
## INDUS    5.547852     374.74201
## CHAS     1.109869      37.14700
## NOX     10.439986     612.85582
## RM      35.572898    5008.10465
## AGE      7.231292     333.01066
## DIS     10.391413     790.14154
## RAD      3.766538      95.57861
## TAX      9.290381     340.77824
## PTRATIO 11.059010    1102.58568
## B        5.342121     226.87366
## LSTAT   26.521693    5206.14390
```



This chuck creates a regression tree analysis of the data predicting median house prices, medv, using all other predictors.

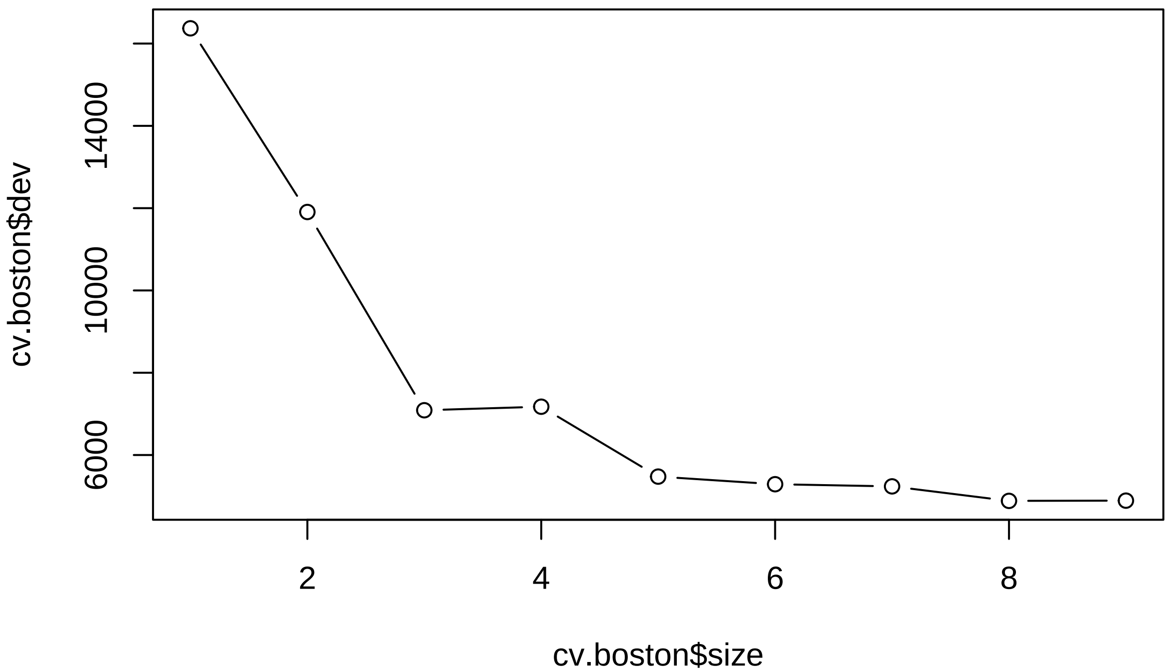
```
##
## Regression tree:
## tree(formula = MEDV ~ ., data = train.data)
## Variables actually used in tree construction:
## [1] "LSTAT" "RM"    "CRIM"  "NOX"
## Number of terminal nodes:  9
## Residual mean deviance:  9.687 = 2364 / 244
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -8.8220 -1.7890 -0.3533  0.0000  1.7110 25.0300
```



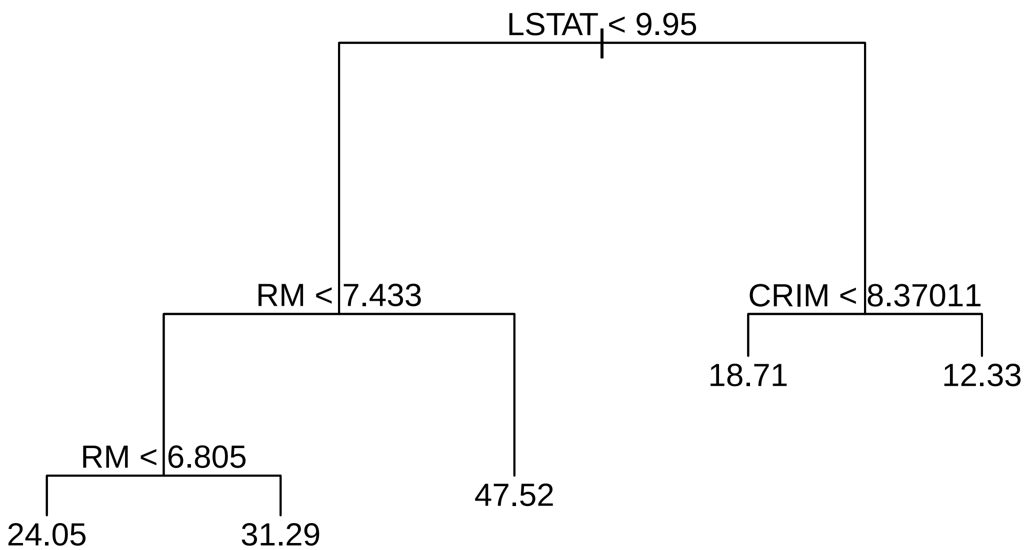
Call predict on the model passing in the test data, assigning it to the variable "tree.pred". then calculate MSE by mean of the square of the difference between the predictions, tree.pred, and true values, true.vals. mean ((pred-true) ˆ 2)

```
## [1] 332.2832
```

Now look for a tree size for pruning. This code creates a plot of candidate tree sizes vs. error (deviation).



Prune the tree using the "prune.tree" function. Use the est.size variable you created above as the value for the "best" parameter. Then, in two statements, call plot and text on the object returned from the prune.tree method to procude a plot of the pruned tree.



Now make predictions and calculate MSE for the pruned tree as you did for the unpruned tree..

```
## [1] 327.1698
```

##Summary and Conclusions

References

https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data (Source [of dataset]: StatLib - Carnegie Mellon University)