Analysis and Prediction of Boston Home Prices

Akanksha Dinesh Jagdale

The University of Texas at Dallas

BUAN 6359.502: Advanced Statistics for Data Science

Dr. Monica Brussolo

5 December, 2020

ABSTRACT

In this paper, I have analyzed and predicted the prices of residential properties in the Boston city using the statistical algorithm of Regression.

I have used the multiple linear regression methodology to model the relationship between the explanatory variables (independent variables) and a response variable (dependent variables) by fitting a linear equation to the observed data. Multiple predictors in the data are used to find the best model for predicting the "MEDV" (median value of house price).

First, I loaded the data into an object data frame and performed summary statistics and Exploratory Data Analysis (EDA) to visually analyze the normality and linearity of the data. Then, I have built a model using regression technique and plotted the residuals vs. fitted values graphs. Looking at the residuals errors, p-value, F-statistics, I have analyzed the performance of the models.

As a part of the result, I found out that there is a significant relationship between the MEDV (price of house) and the RM (no. of rooms), PTRATIO (pupil-teacher ratio) and LSTAT (% of lower status of population) variables.

Keywords: Housing Price Prediction, Multiple Linear Regression

Table of Contents

INTRODUCTION

In today's world, everyone wishes to have a house that suits their lifestyle and provides all the amenities according to the needs. House prices keep on changing very frequently and some of the factors that influence the prices of a house include physical conditions around the neighborhoods, number of rooms in the house, location, other basic local amenities, etc. House price prediction can help real estate agents, estate developers, some government agencies, urban planners, finance professionals and most importantly the homeowners who would make use of such information on a daily basis.

Generally house price predictions are based on the comparisons between the cost price and sales price. But, predicting the prices for future can be difficult as there are a lot many factors that affect the housing market prices. Thus, the development and availability of house price prediction models can prove to be helpful in improving the efficiency of the real estate market. I have attempted to build one such model.

RELATED WORK

Quoting from this paper, (Darshil Shah, 2020, pp. 2-5), paper provides predictions on house prices by using data mining techniques. They have used Decision tree Analysis, Cat Boost algorithm along with Robotic Process Automation for real-time data extraction. Robotic Process Automation involves the use of software robots to automate the tasks of data extraction and the machine learning algorithm is used to predict house prices.

(Bahia, 2013, pp. 1) This project provides real estate market forecasts on home prices by using some of the data mining techniques. They have built a neural network model by using two types of neural networks - First feed forward neural network (FFBP) and Cascade forward neural network (CFBP). The results include comparing the two models to find the best performing model that predicts the prices.

(Hromada, 2015, pp. 1) This paper describes an application that can be used for the analysis of advertisements of real estates published on the Internet in the Czech Republic. The software is used to collect, analyze and assess data about the changes in the real estate market. All the real estate advertisements are stored in a database and are thoroughly analyzed.

(Wittowsky, 2020, pp. 1)This research project has analyzed the housing prices in Dortmund, Germany based on accessibility of good neighborhoods. Walk Score is used for predicting the indicators. The ordinary-least-squares regression (OLS) model is used to characterize the importance of neighborhood factors. The results conclude that other factors like density of supply, walking distances or public transport service quality should also be considered for the prediction.

SYSTEM MODEL

The dataset used for this project can be accessed from Sample Data: Boston Homes | Wolfram Data Repository (wolframcloud.com). The Boston housing data was collected in 1978 and has 506 rows and 14 columns that represent various features for homes in Boston, Massachusetts. All the

input variables are numeric. The missing values in dataset are replaced by computing a statistical value i.e. the mean *[figure-1]*.

Some of the assumptions considered are:

- Normality: The data has a normal distribution *[figure-2]*.
- Linearity: The relationship between the dependent and independent variable is linear i.e. the line of best fit through the data points is a straight line.

PROBLEM STATEMENT

How well can the house prices in Boston City be predicted by using the multiple linear regression algorithm?

SOLUTION

A. Data Description: Below is a brief description of each column in the dataset:

CRIM	Per capita rate by town
ZN	Proportion of residential land zoned for lots over 25000 sq. ft.
INDUS	Proportion of non-retail business acres per town
CHAS	Charles river dummy variable
NOX	Nitric oxide concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner occupied units prior to 1940
DIS	Weighted distances to five employment centers in Boston
RAD	Index of Accessibility to radial highways
TAX	Property tax rate per $10000
PTRATIO	Pupil-teacher ratio
LSTAT	Percentage of lower status by population
MEDV	Median value of owner-occupied homes in $1000

Table1: Data Description

B. Correlation: Correlation measures the interdependence between the variables. A positive correlation indicates that when there is increase in one variable (X), the other variable (Y) also increases and a negative correlation indicates that when there is increase in one variable (X), the other variable (Y) decreases. From the correlation matrix [figure-a], we see that "RM" has a strong positive correlation (0.695) with "MEDV" and "LSTAT" has a strong negative correlation (-0.721) with "MEDV". Features "CHAS" and "DIS" have a weak correlation with "MEDV" *[figure-3]*.

C. Building the Model:

The equation for multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

where,

Y= predicted value of dependent variable

β_0 = y-intercept

β_1 = regression coefficient (β_1) of independent variable (X_1)

ε = model error

(a) Model 1:

```
lm1 <- lm (MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + TAX + PTRATIO + LSTAT, data =
boston_housing)
```

From the plot *[figure-4]*, we see no pattern in the data, and so there is no need for a transformation. Also, from the p-values of the above model *[figure-5]*, we see that features "CRIM", "ZN", "INDUS", "AGE", "TAX" and "NOX" are not significant. So, we build a model excluding those features.

(b) Model 2:

```
lm2 <- lm (MEDV ~ RM + PTRATIO + LSTAT, data = boston_housing)
```

The p-values for all the features are highly significant in this model as seen in *[figure-6]*. After excluding the insignificant features, the F-statistic has also improved from 101.7 to 337.5

However not much improvement is seen in the R-squared value. To improve upon this model, we use a log transformation on the "MEDV" variable.

(c) Model 3:

```
lm3 <- lm (log (MEDV) ~ RM + PTRATIO + LSTAT, data = boston_housing)
```

The output of model 3 shows that the F-statistics is increased to 387.8 *[figure-7]*

The p-values are highly significant and this model looks better than previous models. From the graph plot [figure-e] we can see that the residuals are centered around zero and with similar spread on both the sides.

RESULTS

From the final model, we see that the RM (Average number of rooms per dwelling), PTRATIO (Pupil-teacher ratio) and LSTAT (Percentage of lower status by population) play a significant role in predicting the MEDV (Median value of owner-occupied homes) in the Boston City.

Thus, the final regression equation for the model is:

$$\text{Log (MEDV)} = 3.4645 + 0.1174 * \text{RM} - 0.0396 * \text{PTRATIO} - 0.0343 * \text{LSTAT}$$

Limitations of Linear Regression:

- Linear regression is *sensitive to outliers*.
- *Over fitting* — over fitting means that the regression models random error like noise in the data. Over fitting usually happens when there are too many parameters compared to the number of samples. Therefore, we have only 4 independent variables in our model.
- Linear regression describes a linear relationship between the dependent and independent variables. If there is a nonlinear relationship we compensate by *transforming the target variable* with transformations like log transformation.

CONCLUSION

The goal of this research paper was to determine the attributes that best explained variation in house pricing. Linear regression technique was used to eliminate some of the insignificant predictors and observations. From the final model, we see that the price of houses with more number of room are higher. The house prices are higher in areas with lower pupil-teacher ratios and the prices are lower where the percentage of the lower status of the population is higher.

But, because the data was collected in 1978, it is certain that the factors that will affect the prices today would have changed and it will be interesting to examine these factors. Based on the predicted values, all of these results play an important role when thinking of buying or selling a home.

Future scope for this study can be applying more algorithms and making a few enhancements on the model and the attributes. The accuracy rate can be improved by doing some more transformations of the variables to reduce the error rates. Including some more factors like geographic information for the City of Boston can help predict the prices more accurately. Also the past data in terms of prices, which means considering a larger dataset can be helpful in predicting the prices.

REFERENCES

1. Darshil Shah (2020), House Price Prediction Using Machine Learning and RPA. International Research Journal of Engineering and Technology (IRJET), ISSN: 2395-0056, Volume: 07 Issue: 03 | Mar 2020

2. Bahia, I. S. (2013). A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study. International Journal of Intelligence Science.
3. Hromada, E. (2015). Mapping of Real Estate Prices Using Data Mining Techniques. Procedia Engineering.
4. Dirk Wittowsky, Josje Hoekveld, Janina Welsch & Michael Steier (2020) Residential housing prices: impact of housing characteristics, accessibility and neighboring apartments – a case study of Dortmund, Germany, Urban, Planning and Transport Research.
5. http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

Appendix A

Figure-1: Exploratory Data Analysis

```
#data exploration
dim(boston_housing)
sapply(boston_housing,class)
summary(boston_housing)


#handle missing values
for(i in 1:14)
{
boston_housing[[i]][is.na(boston_housing[[i]])] <-
mean(boston_housing[[i]], na.rm=TRUE)
}
summary(boston_housing)
```

Figure 2: Assumptions

```
#check normality of target variable
ggplot (boston_housing, aes(x=MEDV)) +
geom_histogram (binwidth = 5, color="black", fill="lightblue")
```

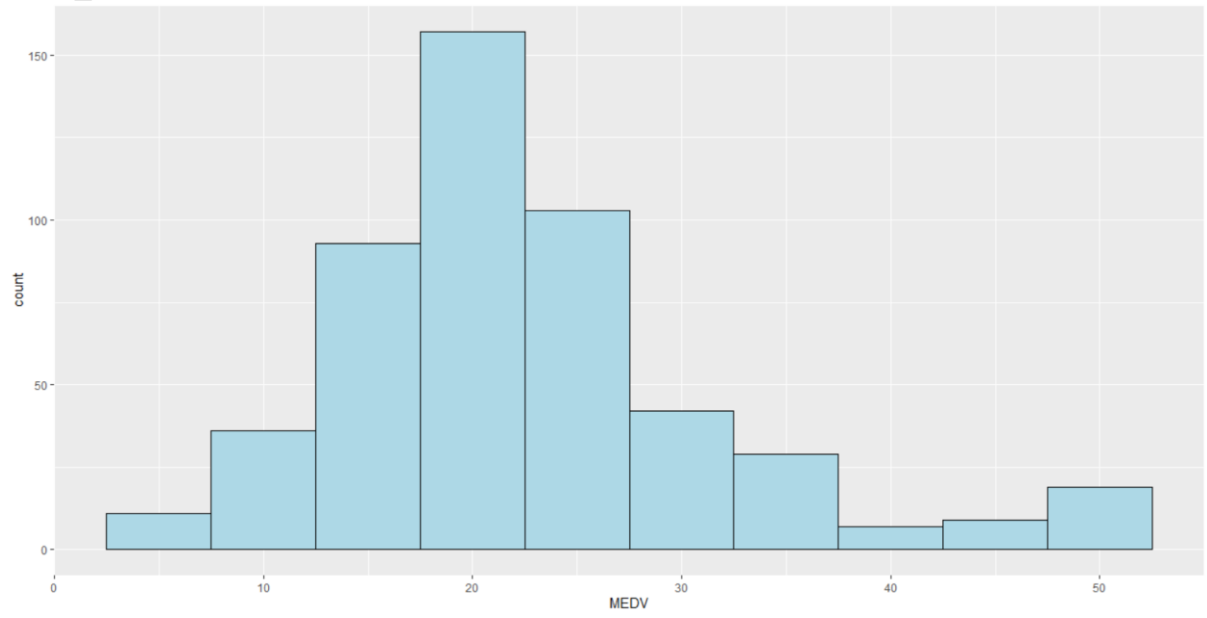
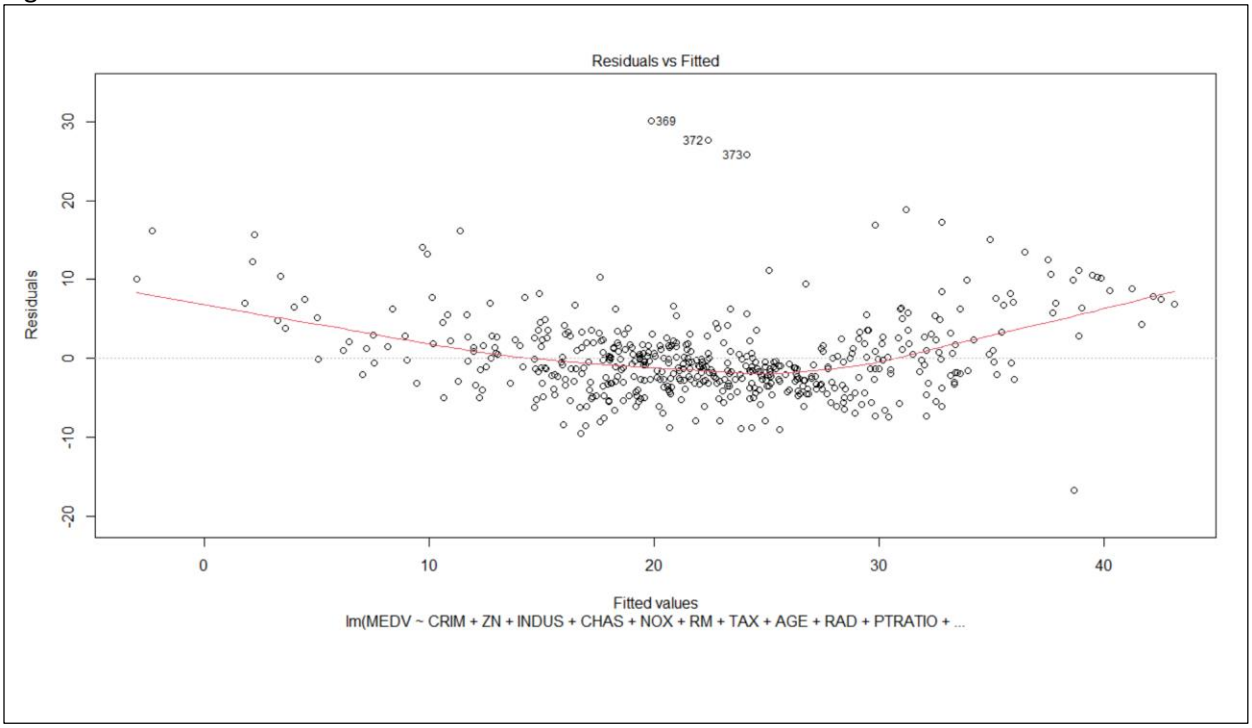


Figure-3: Correlation between variables

```
#check correlation
cor_matrix <- cor(as.matrix(boston_housing,
method=c("Pearson")))
cor_matrix

>cor_matrix

              CRIM          ZN       INDUS         CHAS         NOX         RM         AGE
DIS
CRIM     1.00000000 -0.18292999  0.39116137 -0.052222882  0.41037672 -0.2154338  0.34493361 -
0.36652274
ZN      -0.18292999  1.00000000 -0.51333622 -0.036146534 -0.50228742  0.3165496 -0.54127450
0.63838811
INDUS    0.39116137 -0.51333622  1.00000000  0.058034842  0.74096466 -0.3814574  0.61459225 -
0.69963912
CHAS    -0.05222288 -0.03614653  0.05803484  1.000000000  0.07328555  0.1022839  0.07520637 -
0.09168032
NOX      0.41037672 -0.50228742  0.74096466  0.073285549  1.00000000 -0.3021882  0.71146138 -
0.76923011
RM      -0.21543377  0.31654961 -0.38145737  0.102283891 -0.30218819  1.0000000 -0.24135070
0.20524621
AGE      0.34493361 -0.54127450  0.61459225  0.075206366  0.71146138 -0.2413507  1.00000000 -
0.72435308
DIS     -0.36652274  0.63838811 -0.69963912 -0.091680318 -0.76923011  0.2052462 -0.72435308
1.00000000
RAD      0.60888632 -0.30631636  0.59317646  0.001424954  0.61144056 -0.2098467  0.44998866 -
0.49458793
TAX      0.56652782 -0.30833429  0.71606232 -0.031482822  0.66802320 -0.2920478  0.50058938 -
0.53443158
PTRATIO  0.27338389 -0.40308541  0.38480592 -0.109309953  0.18893268 -0.3555015  0.26272340 -
0.23247054
B       -0.37016342  0.16743135 -0.35459662  0.050054508 -0.38005064  0.1280686 -0.26528227
0.29151167
LSTAT    0.43404449 -0.40754907  0.56735384 -0.046165782  0.57237922 -0.6029620  0.57489289 -
0.48342926
MEDV    -0.37969547  0.36594312 -0.47865733  0.179882500 -0.42732077  0.6953599 -0.38022344
0.24992873

                RAD         TAX    PTRATIO           B       LSTAT       MEDV
CRIM     0.608886320  0.56652782  0.2733839 -0.37016342  0.43404449 -0.3796955
ZN      -0.306316361 -0.30833429 -0.4030854  0.16743135 -0.40754907  0.3659431
INDUS    0.593176456  0.71606232  0.3848059 -0.35459662  0.56735384 -0.4786573
CHAS     0.001424954 -0.03148282 -0.1093100  0.05005451 -0.04616578  0.1798825
NOX      0.611440563  0.66802320  0.1889327 -0.38005064  0.57237922 -0.4273208
RM      -0.209846668 -0.29204783 -0.3555015  0.12806864 -0.60296205  0.6953599
AGE      0.449988663  0.50058938  0.2627234 -0.26528227  0.57489289 -0.3802234
DIS     -0.494587930 -0.53443158 -0.2324705  0.29151167 -0.48342926  0.2499287
RAD      1.000000000  0.91022819  0.4647412 -0.44441282  0.46843967 -0.3816262
TAX      0.910228189  1.00000000  0.4608530 -0.44180801  0.52454474 -0.4685359
PTRATIO  0.464741179  0.46085304  1.0000000 -0.17738330  0.37334313 -0.5077867
B       -0.444412816 -0.44180801 -0.1773833  1.00000000 -0.36888621  0.3334608
LSTAT    0.468439666  0.52454474  0.3733431 -0.36888621  1.00000000 -0.7219746
MEDV    -0.381626231 -0.46853593 -0.5077867  0.33346082 -0.72197464  1.0000000
```

.

Figure-4: Residuals vs. fitted values for model 1



.

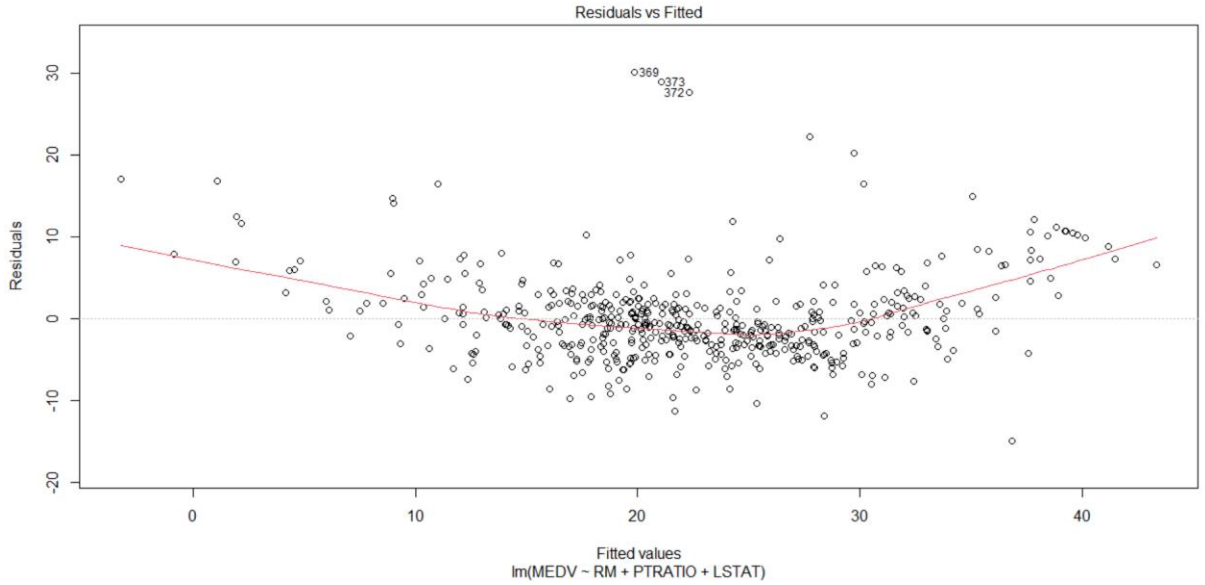
Figure-5: Output for the linear model 1

```
>summary(lm1)

Call:
lm(formula = MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + TAX +
    AGE + PTRATIO + LSTAT, data = boston_housing)
Residuals:
     Min       1Q   Median       3Q      Max
-16.7494  -3.1130  -0.8674   1.9049  30.1550
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.856817   4.858904   5.116 4.48e-07 ***
CRIM        -0.096911   0.034818  -2.783 0.005586 **
ZN           0.001060   0.013661   0.078 0.938209
INDUS        0.061073   0.063255   0.966 0.334757
CHAS         3.355511   0.943393   3.557 0.000411 ***
NOX         -8.592917   3.843593  -2.236 0.025821 *
RM           4.410730   0.436891  10.096  < 2e-16 ***
TAX         -0.011889   0.004047  -2.938 0.003457 **
AGE          0.019159   0.013269   1.444 0.149400
PTRATIO     -0.996349   0.141910  -7.021 7.32e-12 ***
LSTAT       -0.503421   0.052406  -9.606  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5.147 on 494 degrees of freedom
Multiple R-squared:  0.6936,  Adjusted R-squared:  0.6868
F-statistic: 101.7 on 11 and 494 DF,  p-value: < 2.2e-16
```

.

Figure-6: Output for linear model 2

```
>summary(lm2)
Call:
lm(formula = MEDV ~ RM + PTRATIO + LSTAT, data = boston_housing)

Residuals:
    Min      1Q  Median      3Q     Max
-14.943  -3.212  -0.745   1.888  30.149

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.05723    3.95631   4.311 1.95e-05 ***
RM           4.75341    0.42833  11.098  < 2e-16 ***
PTRATIO     -0.94229    0.11954  -7.882 2.01e-14 ***
LSTAT       -0.55109    0.04324 -12.746  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.311 on 502 degrees of freedom
Multiple R-squared:  0.6685,     Adjusted R-squared:  0.6665
F-statistic: 337.5 on 3 and 502 DF,  p-value: < 2.2e-16

>plot2 <- plot(lm2)
>plot2
```



.

Figure-7: Output of the final model

```
>summary(lm3)

Call:
lm(formula = log(MEDV) ~ RM + PTRATIO + LSTAT, data =
boston_housing)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93251 -0.11132 -0.00854  0.11710  0.86290

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.464528   0.167681  20.661  < 2e-16 ***
RM           0.117447   0.018154   6.469 2.34e-10 ***
PTRATIO     -0.039613   0.005067  -7.818 3.17e-14 ***
LSTAT       -0.034371   0.001833 -18.756  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2251 on 502 degrees of freedom
Multiple R-squared:  0.6985,     Adjusted R-squared:  0.6967
F-statistic: 387.8 on 3 and 502 DF,  p-value: < 2.2e-16

plot3 <- plot(lm3)
plot3
```