Finding the Fair Value of a House in Boston

By: Dev Gupta

Table of Contents

Introduction

The Problem

Housing prices have been on the rise particularly in a time plagued by the novel coronavirus. This program has been made to give your house a value based on the facts. It is made to provide information to the common person based on aspects like location, number of rooms, square footage, and other physical attributes. The real-estate business is one that is never ending, but it is pertinent that it be kept in check. Programs like this can do that. While this one does not have perfect accuracy, over time, with more data the accuracy can be greatly improved.

The Data

The dataset I used was one offered by sklearn. This data set was great to use because it had no missing data, and was relatively thorough in considering all aspects. I only wish there were more data points. One can see, as the inputs head more and more towards the outliers, the predictions become less accurate. This only happens at major extremes.

The Elements Broken Down

1. CRIM
    a. Per capita crime rate
2. ZN
    a. Proportion of zoned residential land
3. INDUS
    a. Proportion of zoned non-retail business
4. CHAS
    a. A control variable
5. NOX
    a. Nitric oxide concentration
6. RM
    a. Rooms in the house
7. AGE
    a. Age of house
8. DIS
    a. Distance to city business center
9. RAD
    a. Accessibility to highways
10. TAX
    a. Tax rate
11. PTRATIO
    a. Student-teacher ratio at public schools
12. LSTAT
    a. Percentage of lower financial class within population
13. MEDV

a. Price of the house

Methodology

Importing and Understanding the Data

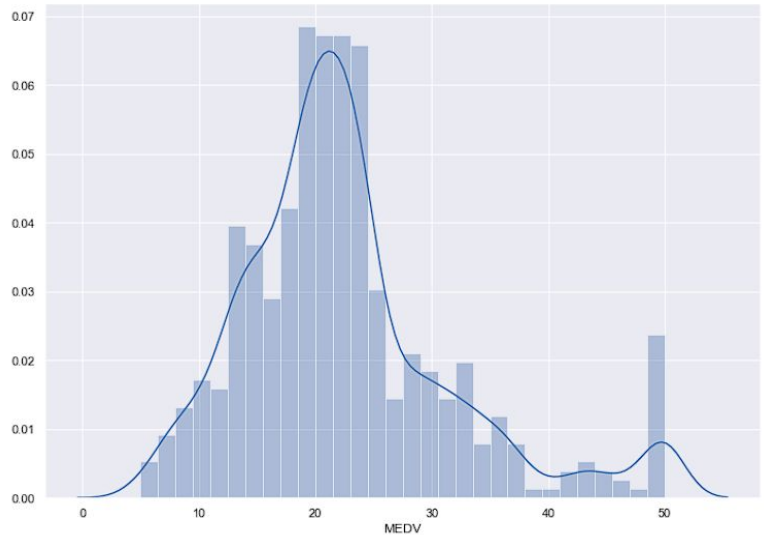
In this section the process of the code will be explained. The first section is straightforward. In the first couple of cells I am solely importing different tools that will be of assistance later. I import Numpy, a mathematical solver, Matplotlib, a plotting tool, Pandas, a data framing tool, and Seaborn, a visualization tool. I will import more as I go on. I then download the dataset from Sklearn. Next I print the keys and redefine them to make a data frame. Below are the first six rows:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33
5	0.02985	0.0	2.18	0.0	0.458	6.430	58.7	6.0622	3.0	222.0	18.7	394.12	5.21
6	0.08829	12.5	7.87	0.0	0.524	6.012	66.6	5.5605	5.0	311.0	15.2	395.60	12.43
7	0.14455	12.5	7.87	0.0	0.524	6.172	96.1	5.9505	5.0	311.0	15.2	396.90	19.15
8	0.21124	12.5	7.87	0.0	0.524	5.631	100.0	6.0821	5.0	311.0	15.2	386.63	29.93
9	0.17004	12.5	7.87	0.0	0.524	6.004	85.9	6.5921	5.0	311.0	15.2	386.71	17.10

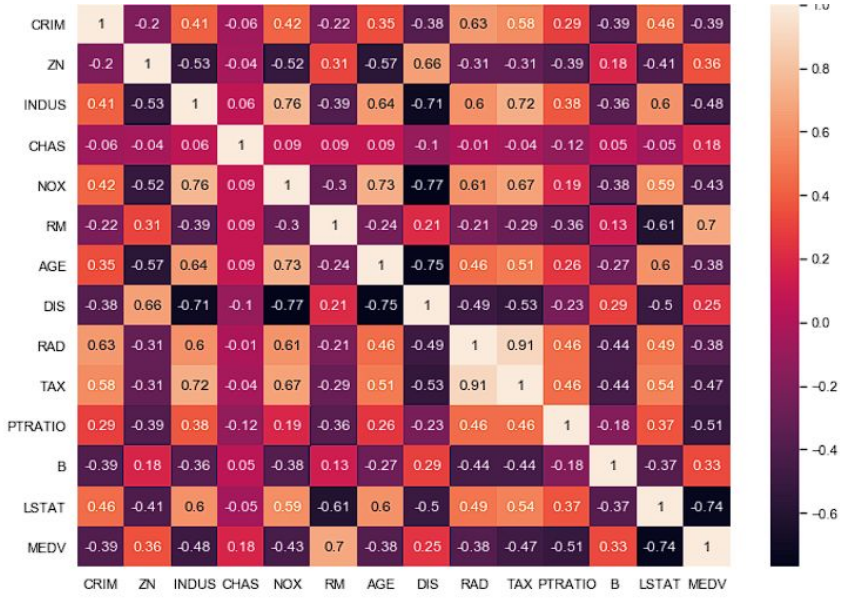
I then add the data I want my regression to predict, the price of the house. One I have added that column this is what the top of the data frame looks like:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

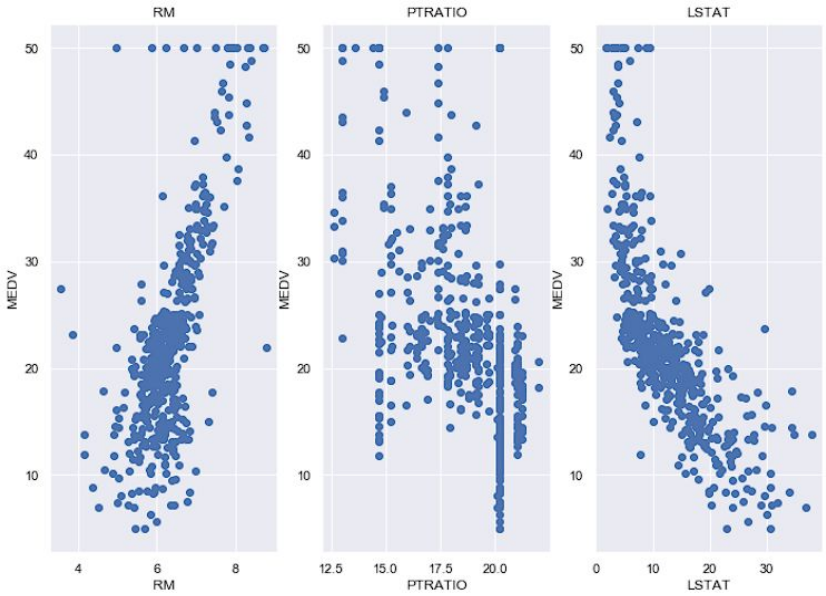
Continuing on, I wanted to see how many houses of different prices there were so I decided to make a histogram. I started off by binning all of the data to make categories. I then plotted it as a histogram, here is the result:



Back to the main project at hand, I next made a heatmap that demonstrated the different correlation between elements. This information would allow me to see what aspects affect price most and make a model based off of that. Here is what the heatmap looked like:



From looking at this heatmap it is clear which elements we should use to predict the MEDV ot the price. It is clear that the elements LSTAT and RM have the greatest correlation. This is the percentage of the lower class population and the amount of rooms per house respectively. Let's show how important this correlation is. Below I have made another form to visualize the correlation. Notice how LSTAT and RM are both very linear compared to PTRATIO which is very spread out.



Preparing and Training the Data

This section is very code heavy and contains very little visual aids. I start by making a split data set. First my data to base the predictions off of, X, this contains both LSTAT and RM. The second is the target, Y, this contains MEDV. I then split the data for training using Sklearn and trained it. I then split the data for testing and saw the results. I will discuss these in the results section.

Testing for Accuracy

In making a regression model, accuracy is very important, so in the coding process, I built in a function that prints the accuracy. I depict it in multiple ways - shown below are two.

```
The model performance for training set
--------------------------------------
RMSE is 5.6371293350711955
R2 score is 0.6300745149331701


The model performance for testing set
--------------------------------------
RMSE is 5.13740078470291
R2 score is 0.6628996975186954
```

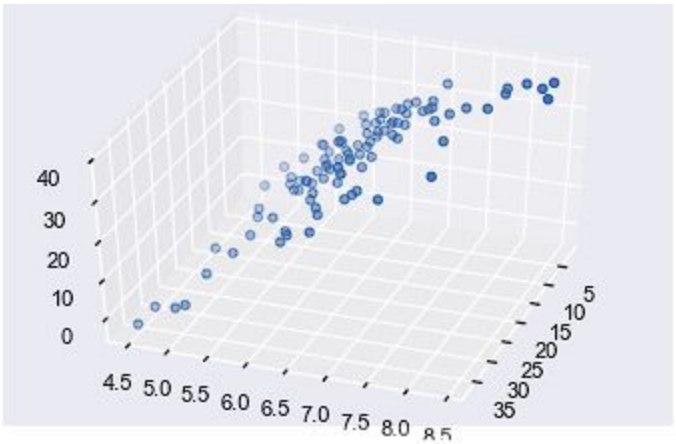
```
Features:
 [[3.13 8.04]]
Prediction: 37.38999403450201
Actual: 37.6

Features:
 [[4.7  6.63]]
Prediction: 29.79290610929409
Actual: 27.9

Features:
 [[8.81  6.417]]
Prediction: 25.867552974668143
Actual: 22.6

Features:
 [[34.77   4.906]]
Prediction: 0.3137082808213769
Actual: 13.8
```

This is a decent accuracy score but it was not as great as I wanted. I tried many methods but this was the optimal score. This made me wonder what was holding back my model, so I looked at the data in another way.



Viewing the data this way shows how confined the data is. The data gets very sporadic towards the extremes and that makes the accuracy drop.

Results

The results of this regression were pretty solid. In the end there was incredible accuracy, at least in the domain covered by the data set, unfortunately once out of that range the results became sporadic and highly inaccurate but that was at an extreme. Built into the program is a basic input-output systems that allows for some user interaction(the house worth is calculated with 4 rooms and a 3% LSTAT:

This is a very sensical answer, for a small house in a pretty well located area, that is a very reasonable price. Of course as mentioned previously there are some flaws, and for discussion of that I would reference you to my methodology section

Discussion

I personally can not make any recommendations based on this report, but the program can, I hope it will one day aid people in getting a fair price for their homes.

Throughout this project I really noticed the importance of a good data set. While this one was usable, it was not ideal. It would have been greatly improved if there was a larger range of data and more data points itself.

Conclusion

This project is aimed at ensuring home buyers are receiving a fair deal. This project utilizes multiple steps to obtain the highest accuracy possible. See the methodology section for more details. To conclude, working on this project has been an eye opening experience. The necessity and vast capabilities of data scientists is astounding. It amazes me that one field can influence so much and give the world so much. It has been a privilege working on not only this project, but this entire certificate as well.