House Price Prediction Using Machine Learning

Akash Singh
Student
School of Computing Science & Engineering
Galgotias University
Greater Noida, India.

Ujjawal chand
Student
School of Computing Science & Engineering
Galgotias University
Greater Noida, India.

Shallendra Kumar
Student
School of Computing Science & Engineering
Galgotias University
Greater Noida, India.

1. Introduction

In this project, we will develop and evaluate the performance and the predictive power of a model trained and tested on data collected from houses in Boston's suburbs.

Once we get a good fit, we will use this model to predict the monetary value of a house located at the Boston's area.

A model like this would be very valuable for a real state agent who could make use of the information provided in a dayly basis.

Getting the Data and Previous Preprocess

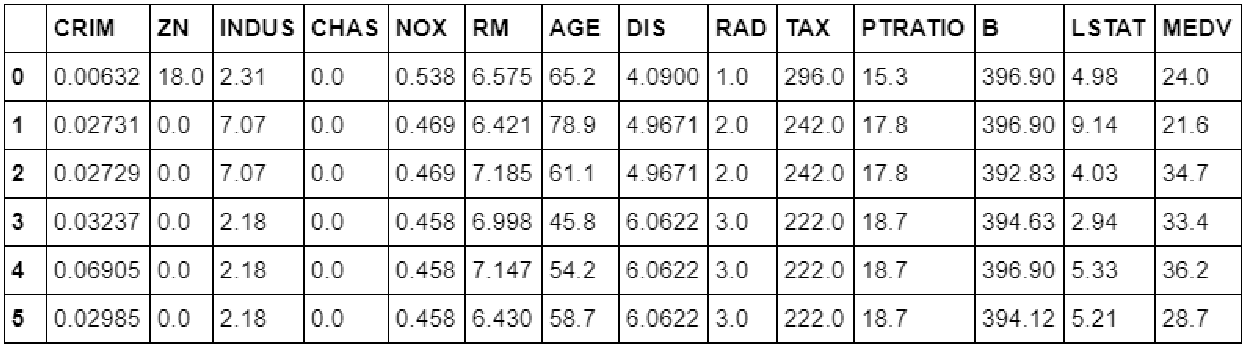
The dataset used in this project comes from the UCI Machine Learning Repository. This data was collected in 1978 and each of the 506 entries represents aggregate information about 14 features of homes from various suburbs located in Boston.

The features can be summarized as follows:

- CRIM: This is the per capita crime rate by town
- ZN: This is the proportion of residential land zoned for lots larger than 25,000 sq.ft.
- INDUS: This is the proportion of non-retail business acres per town.
- CHAS: This is the Charles River dummy variable (this is equal to 1 if tract bounds river; 0 otherwise)
- NOX: This is the nitric oxides concentration (parts per 10 million)
- RM: This is the average number of rooms per dwelling
- AGE: This is the proportion of owner-occupied units built prior to 1940
- DIS: This is the weighted distances to five Boston employment centers

- RAD: This is the index of accessibility to radial highways
- TAX: This is the full-value property-tax rate per $10,000
- PTRATIO: This is the pupil-teacher ratio by town
- B: This is calculated as 1000(Bk — 0.63)², where Bk is the proportion of people of African American descent by town
- LSTAT: This is the percentage lower status of the population
- MEDV: This is the median value of owner-occupied homes in $1000s

This is an overview of the original dataset, with its original features:



	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
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	28.7

For the purpose of the project the dataset has been preprocessed as follows:

- The essential features for the project are: ‘RM’, ‘LSTAT’, ‘PTRATIO’ and ‘MEDV’. The remaining features have been excluded.
- 16 data points with a ‘MEDV’ value of 50.0 have been removed. As they likely contain censored or missing values.
- 1 data point with a ‘RM’ value of 8.78 it is considered an outlier and has been removed for the optimal performance of the model.
- As this data is out of date, the ‘MEDV’ value has been scaled multiplicatively to account for 35 years of markt inflation.

We'll now open a python 3 Jupyter Notebook and execute the following code snippet to load the dataset and remove the non-essential features. Recieving a success message if the actions were correclty performed.

As our goal is to develop a model that has the capacity of predicting the value of houses, we will split the dataset into features and the target variable. And store them in features and prices variables, respectively

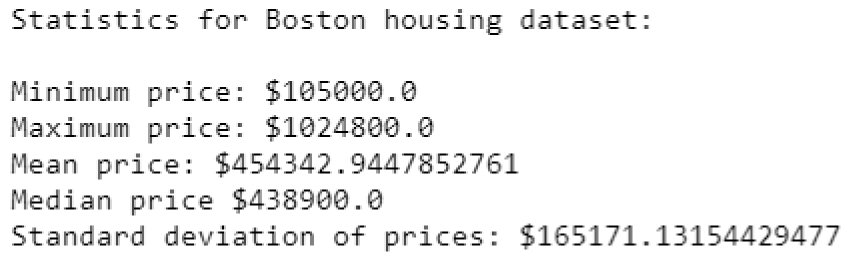
- The features 'RM', 'LSTAT' and 'PTRATIO', give us quantitative information abouth each datapoint. We will store them in *features*.
- The target variable, 'MEDV', will be the variable we seek to predict. We will store it in *prices*.

Data Exploration

In the first section of the project, we will make an exploratory analysis of the dataset and provide some observations.

Calculate Statistics

```
# Minimum price of the data
minimum_price = np.amin(prices)

# Maximum price of the data
maximum_price = np.amax(prices)

# Mean price of the data
mean_price = np.mean(prices)

# Median price of the data
median_price = np.median(prices)

# Standard deviation of prices of the data
std_price = np.std(prices)

# Show the calculated statistics
print("Statistics for Boston housing dataset:\n")
print("Minimum price: ${}".format(minimum_price))
print("Maximum price: ${}".format(maximum_price))
print("Mean price: ${}".format(mean_price))
print("Median price ${}".format(median_price))
print("Standard deviation of prices: ${}".format(std_price))
```



```
Statistics for Boston housing dataset:

Minimum price: $105000.0
Maximum price: $1024800.0
Mean price: $454342.9447852761
Median price $438900.0
Standard deviation of prices: $165171.13154429477
```

Feature Observation

Data Science is the process of making some assumptions and hypothesis on the data, and testing them by performing some tasks. Initially we could make the following intuitive assumptions for each feature:

- Houses with more rooms (higher ‘RM’ value) will worth more. Usually houses with more rooms are bigger and can fit more people, so it is reasonable that they cost more money. They are directly proportional variables.
- Neighborhoods with more lower-class workers (higher ‘LSTAT’ value) will worth less. If the percentage of lower working-class people is higher, it is likely that they have low purchasing power and therefore, they house will cost less. They are inversely proportional variables.
- Neighborhoods with more students to teacher's ratio (higher ‘PTRATIO’ value) will be worth less. If the percentage of students to teacher's ratio people is higher, it is likely that in the neighborhood there are less schools, this could be because there is less tax income which could be because in that neighborhood people earn less money. If people earn less money it is likely that their houses are worth less. They are inversely proportional variables.

We’ll find out if these assumptions are correct through the project.