Boston Housing Dataset Analysis

The following analysis refers to the Boston Housing Dataset, a public dataset that contains information collected by the US Census about house prices in the Boston area. Download made on the site *https://www.kaggle.com/vikrishnan/boston-house-prices*. The machine learning - *multiple linear* regression *method made* in *the* python programming language for *dataset* analysis was used.

The goal of the study is to make a *multiple linear regression model to* try to predict the price of a house in Boston and verify which features *are* the most relevant to the model.

First, the data was loaded into the *jupyter notebook.*

Table 1 - Initial Data

	0.00632	18.00	2.310	0	0.5380	6.5750	65.20	4.0900	1	296.0	15.30	396.90	4.98	24.00
0	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
1	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
2	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
3	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2
4	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222.0	18.7	394.12	5.21	28.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
500	0.06263	0.0	11.93	0	0.573	6.593	69.1	2.4786	1	273.0	21.0	391.99	9.67	22.4
501	0.04527	0.0	11.93	0	0.573	6.120	76.7	2.2875	1	273.0	21.0	396.90	9.08	20.6
502	0.06076	0.0	11.93	0	0.573	6.976	91.0	2.1675	1	273.0	21.0	396.90	5.64	23.9
503	0.10959	0.0	11.93	0	0.573	6.794	89.3	2.3889	1	273.0	21.0	393.45	6.48	22.0
504	0.04741	0.0	11.93	0	0.573	6.030	80.8	2.5050	1	273.0	21.0	396.90	7.88	11.9

As seen in table 1, the data needs preprocessing before applying the model. The *dataset* has 14 attributes. They are:

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's

The task is to predict the value of the property (14) using the rest of the *features*. Attribute 5 will be deleted because it is also a value to be predicted, but it will not be addressed in this analysis.

Tabela 2 – Dados organizados

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRADIO	B-1000	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1.0	273.0	21.0	391.99	9.67	22.4
502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1.0	273.0	21.0	396.90	9.08	20.6
503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1.0	273.0	21.0	396.90	5.64	23.9
504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1.0	273.0	21.0	393.45	6.48	22.0
505	0.04741	0.0	11.93	0.0	0.573	6.030	80.8	2.5050	1.0	273.0	21.0	396.90	7.88	11.9

First, a rearrangement was made in the table, so that we have the correct *names of each feature* in the *dataframe,*as seen in table 2. Because the variable to be predicted is MEDV and NOX was not used in the analysis, they were removed from the dataframe. Through the machine learning *package sklearn,*the data was normalized, except the CHAS feature, because it is a *dummy variable.* After normalization, the data was divided into 2 sets, one to train the model*(x_train,*y_*train)*and another for model testing*(x_test, y_test).* An 80% split was made for testing and 20% for training.

Table 3 - R^2 and Weights Values

R^2	Features	Weights
0,7217	Bias	22,4918
	CRIM	-0,4942
	ZN	0,9715
	INDUS	-0,2791
	CHAS	2,7210
	RM	2,7335
	AGE	-0,0696
	DIS	-2,2094
	RAD	1,8796
	TAX	-2,1014
	PTRADIO	-1,4216
	B-1000	0,9734
	LSTAT	-4,2357

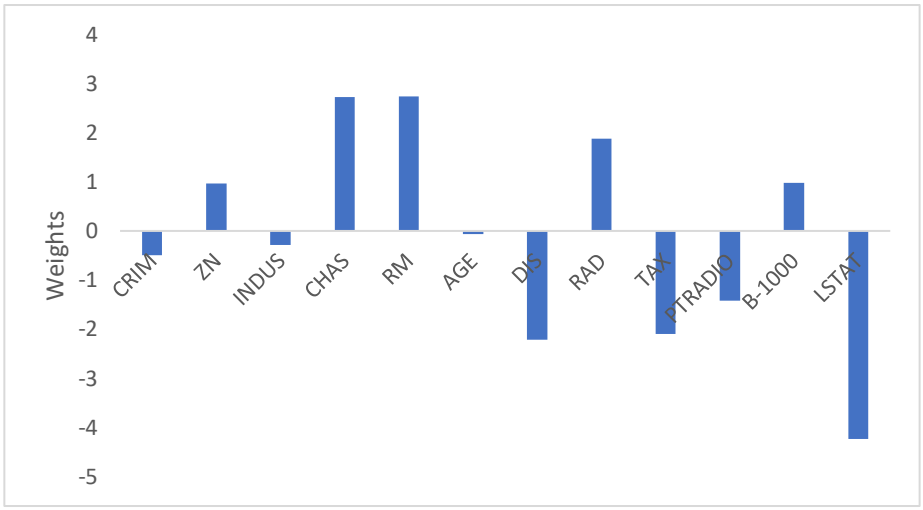

Figure 1 – Weights

With the results obtained, we have a model with approximately 72.1% accuracy. In the evaluation of *weights* values, the more positive, the greater the contribution to the increase in the price of the house and, the more negative, the greater the contribution to the decrease in the price of the house. By table 3, we have that the feature “AGE" is close to 0, that is, almost does not contribute to the final price of the house and therefore can be removed from the analysis. On the weights, the more positive, the more they contribute to an increase in the value of the property and, the lower, to a decrease in the price of the house.

Figure 1 shows that the *features that most contribute* to the increase in the value of the property are, for example, CHAS*,* RM and RAD. These results are expected because, in the case of RM*,* the more rooms in the property, the more expensive it will be. On the other hand, better road infrastructure reflects a higher price of property. Just as the closer the strip of sand a house gets, the higher its value, the same reasoning can be applied in the case of *CHAS*, if the property is close to the river, it will be more valued. For the features with more negative values, we *have LSTAT* that shows the percentage of the population of lower social classes in the Boston area, showing that this influences a lower value of the property. Generally, the lower the level of education of the person, the lower their salary and less conditions to buy a larger and well-located property, so the impact of LSTAT on the model is great. The *DIS feature* shows that the further away from the job centers, the lower the value of the property, which is expected, because the closer to the commercial center of the city, the higher the value of the property. Already, TAX indicates that how much tax is paid for the property, therefore the lower the amount paid in tribute, the cheaper the house. Finally, PTRADIO *is* the relationship between teachers and students in the city, meaning how many students per teacher, so its negative value shows that there are probably many students for each teacher. It may also indicate that the area is not highly valued and concentrated for low-income people, where there are not many schools, mainly public, contributing to the *weight being negative* and indicating lower value housing.

Table 4 - R^2 of the test set

R^2
0,7489

A new test of the model was made to determine the R^2 with the set of and its *obtained* value, seen in table 4, is higher than that *obtained in the training set,* which is not usual, but shows that the model behaved well for new data.