Análise do Boston Housing Dataset

A análise a seguir se refere ao Boston Housing Dataset, um dataset público que contém informações coletadas pelo US Census sobre preços das casas na área de Boston. Download feito no site *https://www.kaggle.com/vikrishnan/boston-house-prices*. Utilizou-se o método de *machine learning - multiple linear regression* feito na linguagem de programação *python* para a análise do *dataset*.

Primeiramente, carregou-se os dados no *jupyter notebook*.

Tabela 1 – Dados Iniciais

	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

Pela tabela 1, os dados precisam de pré-processamento antes do aplicar o modelo. O *dataset* possui 14 atributos. São eles:

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

A tarefa é prever o valor do imóvel (14) utilizando o restante das *features*. O atributo 5 será excluído, pois também é um valor a ser previsto, mas não será abordado nessa análise.

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

Primeiro, foi feito um rearranjo na tabela, para termos os nomes corretos de cada *feature* no *dataframe*, como visto na tabela 2. Como a variável a ser prevista é MEDV e não foi utilizado NOX na análise, elas foram removidas da *dataframe.* Por meio do pacote de *machine learning sklearn*, normalizou-se os dados, exceto o *feature* CHAS, pois ela é uma *dummy variable.* Após a normalização, dividiu-se os dados em 2 sets, um para treinar o modelo (*x_train*, y_*train*) e outro para teste do modelo (*x_test*, *y_test*). Foi feito um split de 80% para teste e 20% para treino.

Tabela 3 – Valores do R^2 e Weights

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

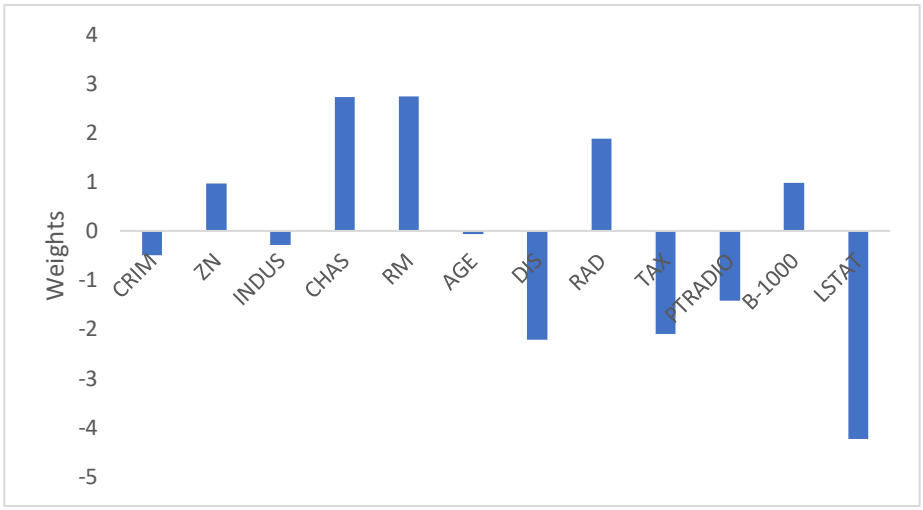

Figura 1 – Weights

Com os resultados obtidos, temos um modelo com, aproximadamente, 72,1 % de precisão. Na avaliação dos valores dos *weights*, quanto mais positivo, maior a contribuição para o aumento do preço da casa e, quanto mais negativo, maior a contribuição para a diminuição do preço da casa. Pela tabela 3, temos que a *feature* "AGE" está próximo de 0, ou seja, quase não contribui para o preço final da casa e, portanto, pode ser retirada da análise. Sobre os *weights*, quanto mais positivos, mais eles contribuem para um aumento no valor do imóvel e, quanto menor, para uma diminuição do preço da casa.

Pela figura 1, foi observado que as *features* que mais contribuem para o aumento do valor do imóvel são, por exemplo, *CHAS*, *RM* e *RAD*. Esses resultados são esperados, porque, no caso de *RM*, quanto mais quartos no imóvel, mais caro ele será. Já RAD, melhor infraestrutura rodoviária reflete num preço maior de imóvel. Assim como quanto mais perto da faixa de areia uma casa fica, maior será seu valor, o mesmo raciocínio pode ser aplicado no caso de *CHAS*, se a propriedade está perto do rio, ela será mais valorizada. Para as *features* com valores mais negativos, temos *LSTAT* que mostra a porcentagem da população de classes sociais mais baixas na área de Boston, mostrando que isso influi para um valor menor do imóvel. Geralmente, quanto menor o nível de educação da pessoa, menor seu salário e menos condições para comprar um imóvel maior e bem localizado, sendo assim o impacto do *LSTAT* no modelo é grande. A *feature DIS* mostra que quanto mais longe dos centros de emprego, menor o valor do imóvel, o que é esperado, pois quanto mais perto do centro comercial da cidade, maior o valor do imóvel. Já, *TAX* indica que o quanto se paga de tributo pela propriedade, portanto quanto menor o valor pago em tributo, mais barata a casa. Por fim, *PTRADIO* é a relação entre professores e alunos na cidade, significando quantos alunos por professor, então seu valor negativo mostra que, provavelmente, há muitos alunos para cada professor. Pode também indicar que a área não é muito valorizada e concentrada de pessoas de baixa renda, onde não há muitas escolas, principalmente públicas, contribuindo para que o *weight* seja negativo e indicando moradias de menor valor.

Tabela 4 – R^2 para set de teste

R^2
0,7489

Foi feito um novo teste do modelo para determinar o R^2 com o *set* de e o seu valor obtido, visto na tabela 4, é maior que o obtido no *set* de treino, o que não é usual, mas mostra que o modelo se comportou bem para novos dados.