



PL10 - RapidMiner: Regressão Linear

Mestrado Integrado em Engenharia Informática

<https://hpeixoto.me/class/dc>

Hugo Peixoto

hpeixoto@di.uminho.pt

2020/2021

Plano de Aula - PL10

 Regressão Linear

 Ficha Exercícios (fe07)



Regressão Linear: Exemplo

Regressão Linear

Contexto e Perspectiva



O **Pedro** está a pensar mudar de casa. Como está à procura do melhor negócio, necessita ter a certeza que o preço que irá pagar é o adequado.

A Pedro sabe que há algumas características importantes quando se procura uma casa e como tal conseguiu reunir um conjunto de dados sobre todas as casas do seu bairro e dos bairros para os quais quer mudar. O objetivo do Pedro é tentar determinar, através desses atributos qual o preço indicado para a casa dos seus sonhos. novos clientes.

O Data Mining pode ajudá-lo a examinar os vários atributos e a influência de cada um no preço da casa.

Adicionalmente conseguirá calcular qual o preço da casa que pretende comprar.

Regressão Linear

Business Understanding

O objetivo do Pedro é bastante claro: determinar qual o valor de mercado de uma determinada habitação.

O Pedro tem um dataset com 120 observações. Ele pretende com esse dataset perceber se o modelo que irá construir é capaz de determinar os valores corretos das casas.

Para atender o objetivo do Pedro, vamos usar um modelo de regressão linear, uma abordagem de modelação estatística que calcula uma relação entre uma resposta escalar (ou variável dependente) e uma ou mais variáveis explicativas (ou variáveis independentes) e que depois usa essa relação para efetuar a previsão.

Regressão Linear

Data Understanding

Os dados do Pedro são:

House_sqft: Tamanho da casa em (sqf)

Num_of_bedrooms: Número de quartos

Num_of_bathrooms: Número de casas de banho

Year_built: ano de construção

Tax_assessed_value: Valor fiscal da casa

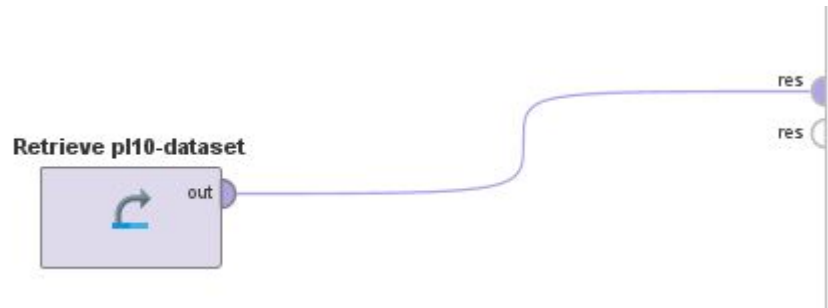
Last_sold_price: Último preço de venda (class)

Regressão Linear

Data Preparation

Download do dataset: pl10-dataset.csv

1. Importe o dataset para o repositório rapidminer (Import Data -> My Computer).
2. Mude para a perspetiva de design e arraste o dataset para a janela do processo.



Data Preparation

Name	Type	Missing	Statistics		
✓ house_sqft	Integer	0	Min 1770	Max 3900	Average 2373.117
✓ num_of_bedrooms	Integer	0	Min 3	Max 5	Average 4.050
✓ num_of_bathrooms	Real	0	Min 2	Max 4	Average 3.021
✓ year_built	Integer	0	Min 1990	Max 2016	Average 2001.800
✓ tax_assessed_value	Integer	0	Min 195000	Max 445000	Average 292783.333
✓ last_sold_price	Integer	0	Min 196358	Max 450842	Average 294614.475

Data Preparation


3. A regressão linear é um modelo preditivo e, portanto, precisa de um atributo para ser designado como label - este é o atributo alvo, aquilo que se pretende prever.

O dataset importado não tem uma definição inicial de qual é o nosso label. Isto é, o alvo do nosso modelo. Para tal temos de utilizar o operador “Set Role” para definir. Neste caso, o nosso label é o “last_sold_price”.

Retrieve pl10-dataset




Parameters ✕

 **Set Role**

attribute name ⓘ

target role ⓘ

set additional roles  Edit List (0)... ⓘ

Data Preparation

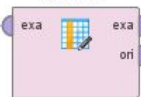
4. O próximo passo é definir os dois conjuntos de treino e teste. Para tal usamos o operador “split data”.

Neste operador podemos definir as percentagens a utilizar, neste caso usando as definições base. Para a definição das percentagens vamos usar 70% para treinar, 30% para testar.

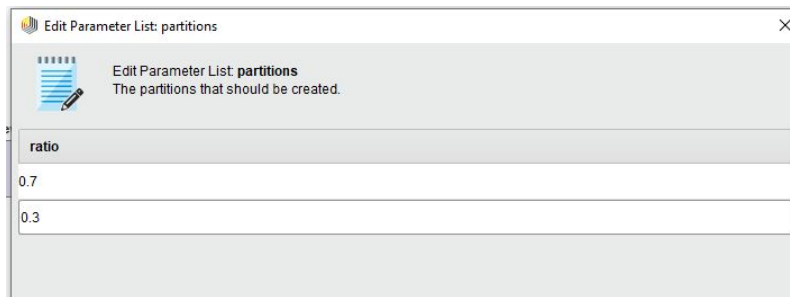
Retrieve p10-dataset



Set Role

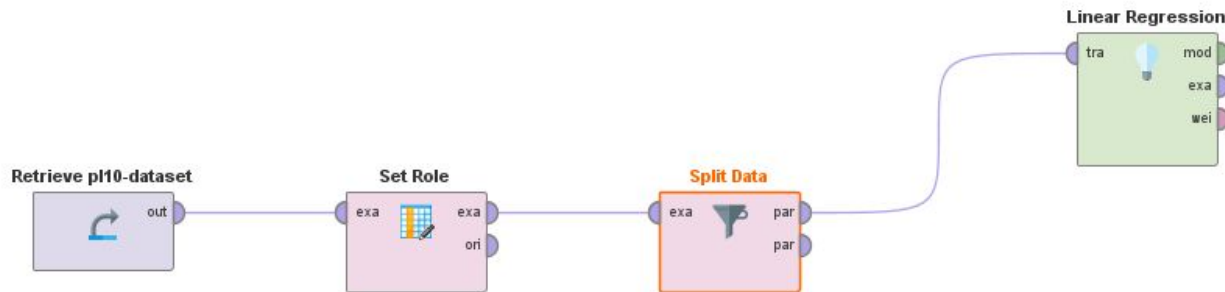


Split Data



Modeling

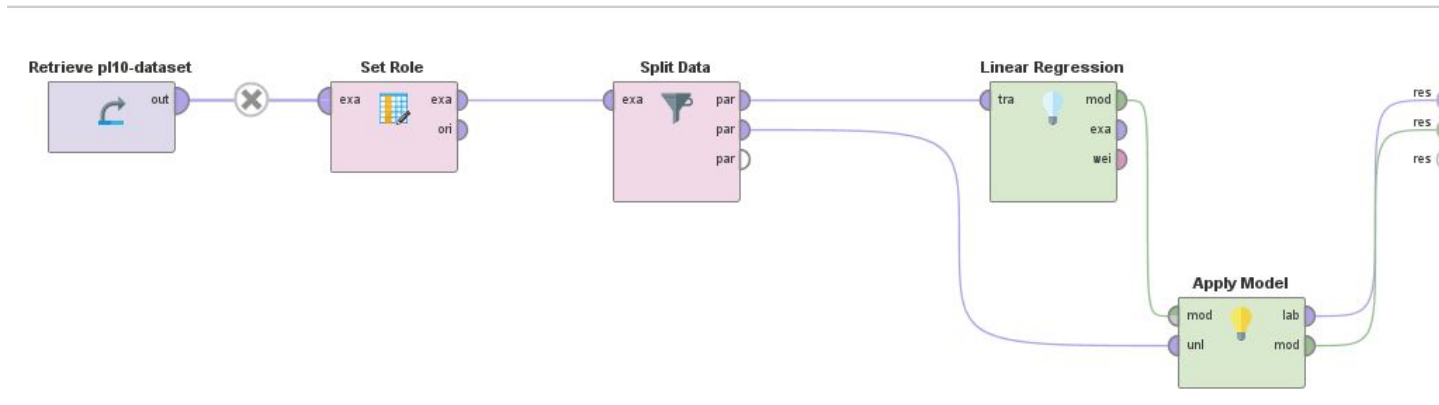
1. Encontrar o operador 'Linear Regression' e arraste-o para a janela do processo. Associe este operador ao fluxo de treino, como mostrado na figura abaixo.





Modeling

- O passo seguinte na modelação é usar um operador do tipo 'Apply Model' para ligar o fluxo de treino ao fluxo de teste. Procure este operador e arraste-o para a janela do processo. Certifique-se de conectar as portas lab e mod às portas res como ilustrado na figura. De seguida devemos colocar também a criação do modelo como modelo de saída para procedermos depois à avaliação.



Evaluation

Corra o modelo. O facto de existirem duas saídas do operador 'Apply Model' conectadas às portas res, resultará em dois separadores na perspectiva de resultados. Vamos examinar primeiro o separador LinearRegression.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
house_sqft	1.737	0.795	0.016	0.580	2.186	0.032	**
num_of_bathrooms	-1192.239	660.564	-0.009	0.683	-1.805	0.075	*
year_built	44.031	78.159	0.006	0.223	0.563	0.575	
tax_assessed_value	0.996	0.012	0.989	0.062	79.850	0	****
(Intercept)	-85701.545	153573.263	?	?	-0.558	0.578	

LinearRegression

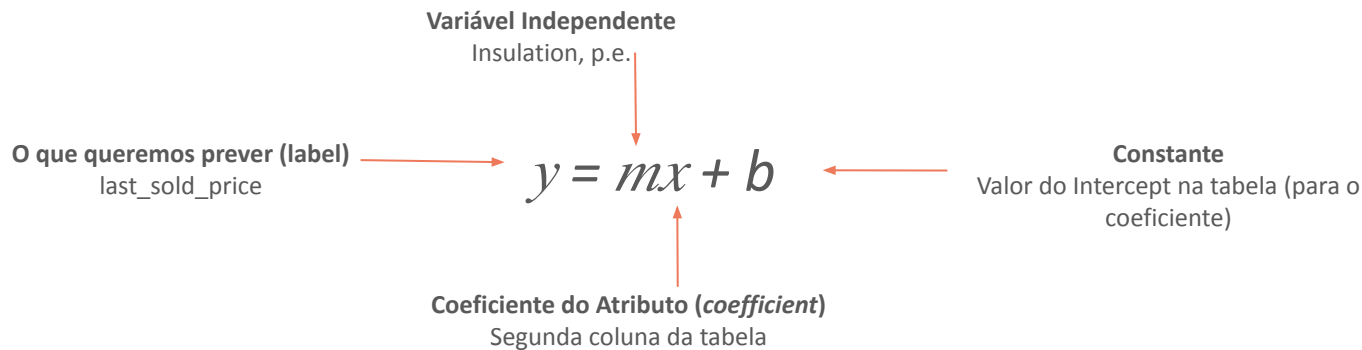
Coefficiente: Valor positivo tem impacto positivo no valor; Valor negativo
Tem impacto negativo.

p-Value: < 0.05 Define quão importante

```
1.737 * house_sqft
- 1192.239 * num_of_bathrooms
+ 44.031 * year_built
+ 0.996 * tax_assessed_value
- 85701.545
```

Evaluation

A modelação de regressão linear tem como objetivo determinar a proximidade de uma determinada observação com uma linha imaginária que representa a média ou o centro de todos os pontos no conjunto de dados.



Se tivéssemos uma casa com uma área de 1900 sqf, a nossa fórmula usando esses valores da área seria

$$y = 1900 \times 1.737 + (- 85701.545)$$

Regressão Linear

Evaluation

- ? Como podemos configurar esta fórmula linear quando temos várias variáveis independentes?
- ? O resultado do operador LinearRegression possui apenas quatro variáveis. O que aconteceu com `number_of_bedrooms`?

Regressão Linear

Evaluation

? Como podemos configurar esta fórmula linear quando temos várias variáveis independentes?

$$y = mx + mx + mx \dots + b$$

Por exemplo:

- house_sqf: 1770
- num_of_bathrooms: 2
- year_built: 1990
- tax_assessed_value: 195000

$$y = 1.737 * 1770 - 1192.239 * 2 + 44.031 * 1990 + 0.996 * 195000 - 85701.545$$

A previsão para o last_sold_price desta casa seria de $(y) = 196\,750.27$, aproximadamente 196 750.

Regressão Linear

Evaluation

? O resultado do operador LinearRegression possui apenas quatro variáveis. O que aconteceu com number_of_bedrooms?

O number_of_bedrooms não era uma variável estatisticamente significativa para prever o preço da casa neste dataset e, portanto, foi removido pelo RapidMiner.

Quando o RapidMiner avaliou a influência que cada atributo no dataset exercia sobre o preço de cada residência representada no dataset de treino, o número de quartos era tão pouco influente que o seu peso na fórmula foi definido como **zero**.

Evaluation

Ainda nos resultados, no separador ExampleSet, selecione a opção Data View. Podemos observar que o RapidMiner previu o custo de cada casapara que o Pedro possa perceber como se deve movimentar no mercado.

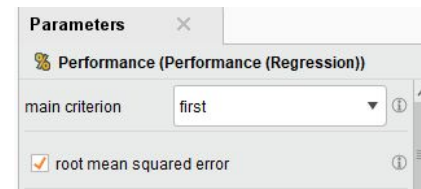
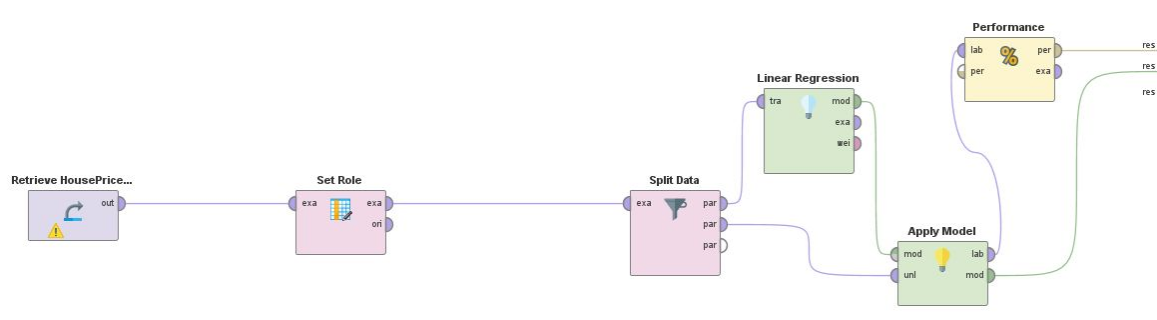
Row No.	last_sold_pr...	prediction(la...	house_sqft	num_of_bed...	num_of_bat...	year_built	tax_assess...
1	197715	196750.274	1770	3	2	1990	195000
2	197816	196750.274	1770	3	2	1990	195000
3	207027	206249.012	1850	3	2.500	1990	205000
4	210519	206249.012	1850	3	2.500	1990	205000
5	211274	213218.131	1850	3	2.500	1990	212000
6	212560	213304.989	1900	4	2.500	1990	212000
7	224194	223392.365	1925	4	2.500	1992	222000
8	232955	230877.343	1992	4	3	1992	230000
9	236792	236019.249	1985	4	3	1996	235000
10	241453	241085.254	1985	4	3	1998	240000
11	241882	241085.254	1985	4	3	1998	240000
12	248234	246306.398	2125	4	3	1998	245000

Evaluation

O Pedro tem agora uma previsão para as casas, mas como será que o modelo se comportou? Será que o modelo estará à altura da exigência do Pedro para garantir que os preços estão a ser bem calculados?

Como validar esse comportamento?

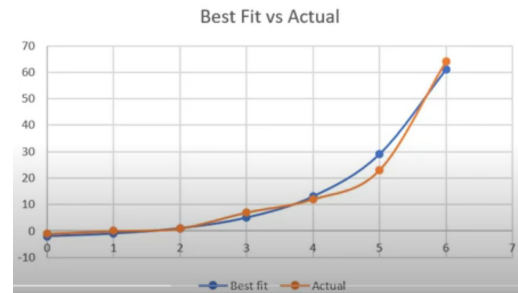
1) Volte para o separador Design e procure o operador “Performance” e adicione-o. O que vamos avaliar é o “root mean squared error”.



Evaluation

Root Mean Square Error (RMSE): Desvio padrão entre os valores previstos e os valores observados.

`root_mean_squared_error: 1968.844 +/- 0.000`



Regressão Linear

Resumo

A regressão linear é um modelo preditivo que usa datasets de treino e scoring para gerar previsões numéricas. É importante lembrar que a regressão linear usa dados numéricos para todos os seus atributos.

Cada atributo no dataset é avaliado estatisticamente pela sua capacidade de prever o atributo do tipo label. Os atributos com fraca capacidade de previsão são removidos do modelo.

Depois das previsões de regressão linear serem calculadas, os resultados podem ser resumidos para determinar se há diferenças nas previsões em subconjuntos dos dados de scoring. À medida que mais dados são recolhidos, estes podem ser adicionados dataset de treino para o tornar mais robusto ou expandir os intervalos de alguns atributos para incluir ainda mais valores.

É muito importante lembrar que os intervalos para os atributos de scoring devem estar dentro dos intervalos dos atributos de treino para garantir previsões válidas.



Ficha de Exercícios 07



PL10 - RapidMiner: Regressão Linear

Mestrado Integrado em Engenharia Informática

<https://hpeixoto.me/class/dc>

Hugo Peixoto

hpeixoto@di.uminho.pt

2020/2021