Ficha 8 - DC

João Nunes (A82300)
Luís Braga (A82088)

03/04/2020

Conteúdo

1 Parte I

1.1 Que tipo de dados a regressão linear espera para todos os atributos? Qual o tipo de dados do atributo previsto quando este for calculado?

A regressão linear necessita de dados numéricos para conseguir prever o *target value*. Portanto, quando o atributo é calculado como recebe dados numéricos irá também produzir um dado numérico.

1.2 Porque é que os intervalos de atributos são tão importantes ao realizar data mining através de regressão linear?

Os intervalos de atributos são importantes uma vez que ao executar o modelo, é necessário verificar que nenhum dos atributos de teste possui valores fora dos atributos de treino, caso contrário seria impossibilitado o uso do modelo de regressão.

1.3 O que são coeficientes de regressão linear? O que significa 'peso', neste contexto?

Os coeficientes de regressão linear indicam que com o aumento dessa variável as suas variáveis dependentes também irão aumentar. Com o coeficiente negativo espera-se um comportamento contrário ao relatado anteriormente. O "peso"deste atributo neste contexto significa a importância dada ao atributo.

1.4 Qual é a fórmula matemática de regressão linear e como é organizada?

A fórmula matemática de regressão linear segue um comportamento linear, ou seja, $y = mx + b$, onde o y representa a variável a prever, o m representa a variável independente, o x é o coeficiente desse atributo e o b é a constante determinada pelos cálculos do modelo, e corresponde ao *intercept*.

1.5 Como é que resultados da regressão linear são interpretados?

Os resultados podem ser resumidos para determinar se há diferenças nas previsões em subconjuntos dos dados de *teste*. A partir dos resultados da regressão linear é possível determinar uma fórmula geral que permite prever para um dado atributo do *dataset* quanto do *target* é que é preciso.

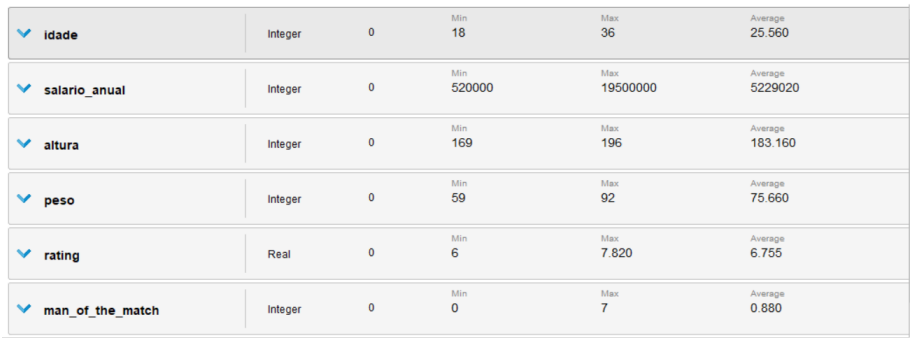
2 Parte II

2.1 Selecione uma organização desportiva profissional de que goste ou que conheça. Localize o site da organização e pesquise estatísticas, factos e números sobre os atletas dessa organização.Crie um dataset (usando o Excel por exemplo) e defina alguns atributos (pelo menos três ou quatro) para armazenar dados sobre cada atleta. Alguns atributos possíveis que pode considerar podem ser o salário anual,pontos_por_jogo, anos_como_pro, altura, peso, idade etc. A lista é potencialmente ilimitada, variará de acordo com o tipo de desporto que escolher e dependerá dos dados disponíveis. O objetivo deste exercício será prever o salário dos atletas, portanto este deve ser um atributo obrigatório. PS:Lembre-se que a regressão linear só trabalha com dados numéricos.

Portanto, de modo a construir um *dataset* futebolístico, foram retirados dados para uma folha de cálculo *excel* de três fontes diferentes. A primeira foi o spotrac, de onde foi possível extrair o salário anual em libras, de seguida foi possível obter dados físicos dos jogadores (altura e peso) através do *website* zerozero, por fim foram retiradas estatísticas acerca dos jogadores, como *rating* médio por jogo e número de vezes que o dado jogador foi considerado homem do jogo através do *website* whoscored.

2.2 Pesquise as estatísticas de cada um dos atributos que selecionou e insira-as como observações na sua folha. Tente encontrar o maior número possível –pelo menos 40, a fim de atingir pelo menos um nível básico de validade estatística. Quanto mais melhor.Divida as observações do seu datas em duas partes: uma parte de treino e uma parte de scoring. Certifique-se que tem pelo menos 20 observações no dataset de treino e pelo menos 20 no dataset de scoring. Como vamos tentar prever o salário dos atletas do dataset de scoring, não precisa de procurar nem preencher a coluna do salário para estes atletas. Guarde FE08 dois ficheiros CSV (treino e scoring), como nomes distintos, carregue-os no RapidMiner e arreste-os para um novo processo.

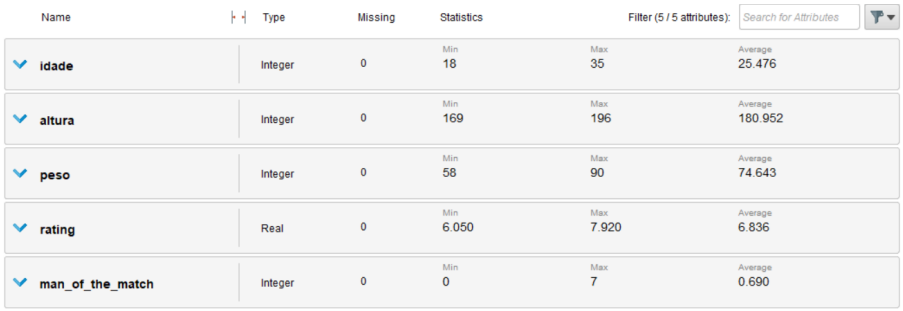
Portanto, tal como foi pedido os dados foram divididos em dois *datasets*, um *dataset* para testes e outro para *treino*. O *dataset* de treino possui o conjunto total dos atributos, ou seja:



idade	Integer	0	Min 18	Max 36	Average 25.560
salario_anual	Integer	0	Min 520000	Max 19500000	Average 5229020
altura	Integer	0	Min 169	Max 196	Average 183.160
peso	Integer	0	Min 59	Max 92	Average 75.660
rating	Real	0	Min 6	Max 7.820	Average 6.755
man_of_the_match	Integer	0	Min 0	Max 7	Average 0.880

Figura 1: Atributos do dataset de treino.

Estes dados para o *dataset* de treino foram retirados de duas equipas da *Premier League*, o *Manchester United* e o *Liverpool*. Por sua vez os atributos para o *dataset* de teste foram retirados de duas outras equipas o *Manchester City* e o *Tottenham*.



Name	Type	Missing	Statistics		Filter (5 / 5 attributes): Search for Attributes
idade	Integer	0	Min 18	Max 35	Average 25.476
altura	Integer	0	Min 169	Max 196	Average 180.952
peso	Integer	0	Min 58	Max 90	Average 74.643
rating	Real	0	Min 6.050	Max 7.920	Average 6.836
man_of_the_match	Integer	0	Min 0	Max 7	Average 0.690

Figura 2: Atributos do dataset de teste.

Estas quatro equipas foram escolhidas em específico devido à semelhança dos valores pagos em salários aos jogadores, de modo a ser justo na escolha dos valores.

2.3 Repita os passos no RapidMiner tal como descritos nos slides da aula e após executar o seu modelo, na secção dos resultados, examine os coeficientes dos atributos e as previsões para os salários dos atletas no conjunto de scoring.

Portanto, e como é possível observar existem algumas discrepâncias no *range* dos valores em ambos os *datasets*. Existem discrepâncias no atributo *peso*, *rating* (embora seja mínima) e *idade*. Como tal e na tentativa de minimizar a diferença de *range* dos valores de ambos atributos foi colocado, por exemplo para o peso, um filtro para os valores da peso superiores a 59 no *dataset* de teste e um outro filtro para valores de peso inferiores a 90 no *dataset* de treino.

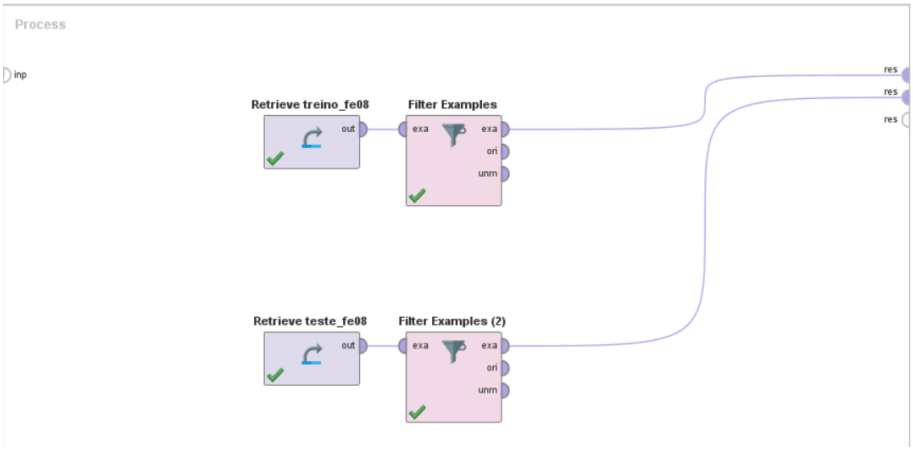


Figura 3: Modelo inicial.

Contudo, mesmo após aplicar este filtro foi possível observar que em ambos *datasets* no máximo o peso ficou limitado a *90*, contudo no caso do *dataset* de teste após aplicar o filtro inferiormente ficou limitado a 60 (segundo menor valor no *dataset*), o que se deve devido a uma falta de valores. Nos outros casos os filtros também não surtem efeito para limitar o *range* de valores, devido a novamente não haver valores suficientes disponíveis no *dataset*.

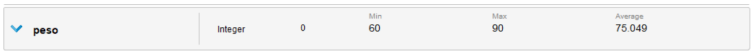


Figura 4: Range de valores peso dataset de teste após filtro.

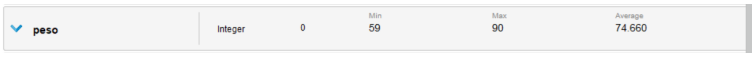


Figura 5: Range de valores peso dataset de treino após filtro.

Após passar esta fase inicial de tratamento dos dados, procedeu-se à construção do modelo de regressão linear, tendo sido seguidos os passos dos *slides* das aulas. No final foi possível obter o seguinte modelo de regressão linear.

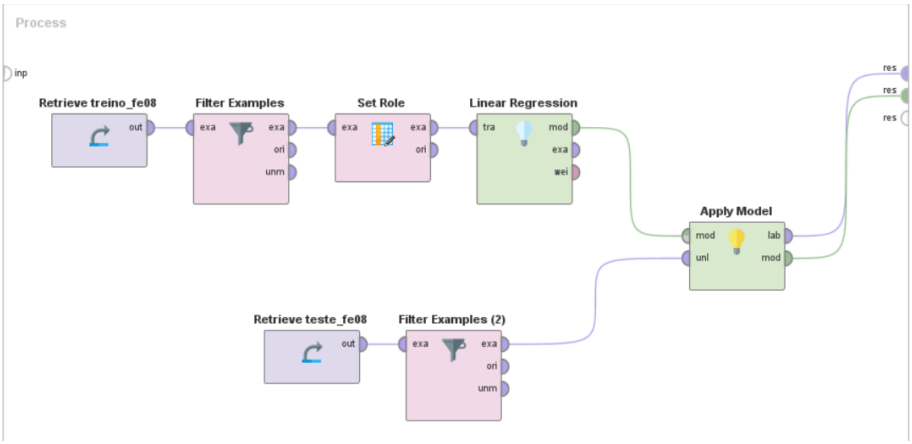


Figura 6: Modelo de regressão linear.

2.4 Relate seus resultados:

2.4.1 (a) Que atributos têm maior peso?

Após correr o modelo anterior foi possível verificar que o atributo *rating* é o que possui maior valor no coeficiente, como tal o *rating* tem maior peso no que toca ao vencimento do jogador, de seguida é a idade e por fim o peso do jogador.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
idade	324871.651	136991.668	0.316	0.978	2.371	0.022	**
peso	48306.994	64860.548	0.098	1.000	0.745	0.460	
rating	3878843.270	1138561.862	0.454	0.966	3.407	0.001	***
(Intercept)	-32779324.545	10319891.241	?	?	-3.176	0.003	***

Figura 7: Resultados da aplicação da regressão linear.

2.4.2 (b) Algum atributo foi removido do conjunto de dados por não ter uma boa capacidade de previsão? Em caso afirmativo, quais e por que você acha que eles não eram eficazes na previsão?

Sim, foram removidos dois atributos do conjunto de dados, foi removido o atributo *man_of_the_match*, foi também removido o atributo *altura*. O primeiro atributo poderá ter sido removido devido a existir pouca variabilidade e um número elevado de valores a zero. O segundo atributo poderá ter sido removido por simplesmente não existir correlação entre a altura e o vencimento recebido pelo jogador.

2.4.3 (c) Procure alguns dos salários de alguns dos seus atletas nos dados de scoring e compare o salário real com o previsto. Está perto?

No que toca à previsão efetuada pela regressão linear, foi possível verificar que para os 10 primeiros jogadores, que pertencem ao *Manchester City*, ordenados por ordem decrescente de salário, foi possível obter a seguinte previsão:

Row No.	prediction(s...	idade	altura	peso	rating	man_of_the...
1	9997524.354	27	181	68	7.920	7
2	6239660.809	24	170	69	7.190	0
3	8275986.141	30	173	70	7.200	2
4	8911737.947	33	173	67	7.150	1
5	8077429.551	33	179	69	6.910	0
6	5503394.773	24	173	65	7.050	1
7	5898372.120	28	180	80	6.630	0
8	5946389.407	22	191	82	7.120	1
9	6440264.792	24	191	86	7.030	0
10	8994070.731	28	179	67	7.590	3

Figura 8: Previsões para os 10 jogadores com maior salário no Manchester City.

O valor real que os jogadores do *Manchester City* recebem poderá ser consultado na seguinte tabela.

Kevin De Bruyne	M	27	£18,200,000	£350,000
Raheem Sterling	M	24	£15,600,000	£300,000
Sergio Aguero	F	30	£11,967,000	£230,135
David Silva	M	33	£11,440,000	£220,000
Fernandinho Luis Roza	M	33	£7,800,000	£150,000
Bernardo Silva	M	24	£7,800,000	£150,000
Ilkay Gundogan	M	28	£7,280,000	£140,000
Rodrigo Hernández Cascante	M	22	£6,300,000	£121,154
Aymeric Laporte	D	24	£6,240,000	£120,000
Riyad Mahrez	F	28	£6,240,000	£120,000

Figura 9: Vencimento real dos jogadores do Manchester City.

Comparando os valores previstos com os valores reais é possível verificar que para os primeiros casos (4 primeiros) a previsão ficou sempre abaixo do valor real recebido pelos jogadores. A partir da quinta entrada o valor previsto aproxima-se mais um bocado do real, sendo que em em algumas situações (entrada número 5, 9 e 8) o valor previsto assemelha-se ao valor real.

No que toca à previsão efetuada pela regressão linear para os jogadores do *Tottenham*.

Row No.	prediction(s...	idade	altura	peso	rating	man_of_the...
22	7522582.159	25	188	80	7.300	3
23	4260163.862	22	181	76	6.760	1
24	7673262.955	26	183	78	7.280	3
25	5035932.517	22	188	76	6.960	1
26	8788184.465	32	188	80	7.040	1
27	8064481.554	31	189	87	6.850	0
28	7651800.290	30	187	90	6.790	0
29	5670828.644	27	184	78	6.680	0
30	5870672.113	26	172	72	6.890	0
31	7188764.639	26	176	76	7.180	2
32	2932552.774	22	177	71	6.480	0

Figura 10: Previsões para os 10 jogadores com maior salário no Tottenham.

O valor real recebido pelos jogadores segue-se na seguinte tabela.

Harry Kane	F	25	£10,400,000	£200,000
Tanguy Ndombele	M	22	£10,400,000	£200,000
Heung-Min Son	F	26	£7,280,000	£140,000
Dele Alli	M	22	£5,200,000	£100,000
Hugo Lloris	GK	32	£5,200,000	£100,000
Jan Vertonghen	D	31	£5,200,000	£100,000
Toby Alderweireld	D	30	£4,160,000	£80,000
Erik Lamela	M	27	£4,160,000	£80,000
Lucas Moura	M	26	£4,160,000	£80,000
Serge Aurier	D	26	£3,640,000	£70,000

Figura 11: Vencimento real dos jogadores do Tottenham.

No caso do *Tottenham*, mais uma vez os dois primeiros salários previstos ficaram abaixo do real, no caso da terceira, quarta e décima entrada os salários ficaram próximos do valor real recebido. Nos restantes casos o salário previsto ficou sempre acima do real.

2.4.4 (d) Que outros atributos acha que ajudariam o seu modelo a prever melhor os salários dos atletas profissionais?

No que toca a melhorias, primeiramente seria interessante contabilizar outros atributos como o número de jogos jogados como profissional, o número de anos que o jogador é profissional, a média de golos, a média de duelos ganhos (no caso dos defesas) e entre outras estatísticas.

De modo a melhorar a previsão também seria necessário recolher muitos mais dados (é um processo muito demorado), uma vez que algoritmos *ML* são *data hungry*, ou seja, para funcionarem corretamente necessitam de muitos dados, muitos mais que as 42 entradas no *dataset* de teste e as 50 no *dataset* de treino.

Para além disso também seria melhor conseguir angariar dados de uma fonte heterogénea e de confiança de modo a manter os dados consistentes, claros e concisos.