my first regression project

Carlos Soares

U. PORTO

FEUP **FACULDADE DE ENGENHARIA**
UNIVERSIDADE DO PORTO
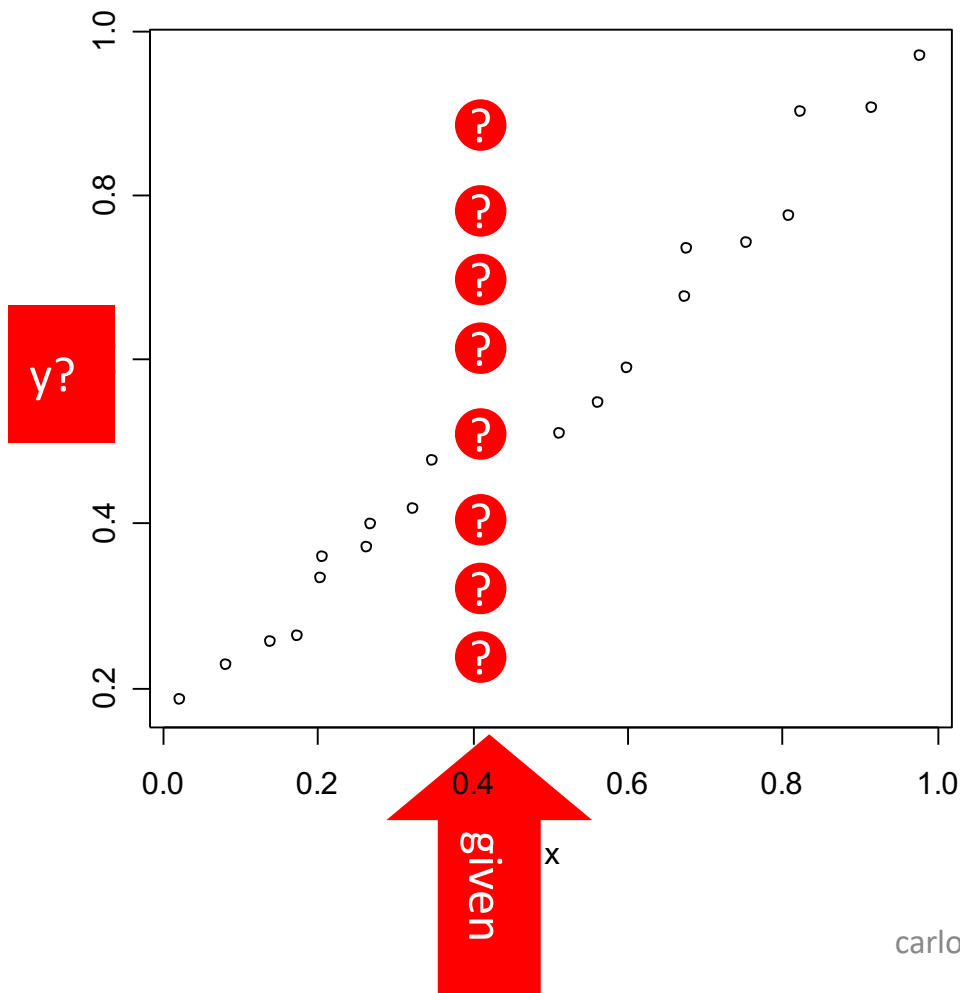
predictive: regression to estimate customer value

- y is customer value
- x is family income
  - and other characteristics
  - ... only 1 here for simplicity

plan & goals

- regression
  - introduction
  - my first regression (in RM...)
- linear regression
- evaluation of Regressors
- other algorithm

- regression concepts
  - interpretation of the linear model
  - evaluation measures of regression models
- understand the need to use different sets of data for modelling and for evaluation
- know how to evaluate the results of a classification model
  - conceptually and in RapidMiner

REGRESSÃO

previsão de valores numéricos: abordagem de análise de dados

- novas observações para as quais queremos fazer a previsão
  - ex. nova zona da cidade

- observações conhecidas
  - ex. bairros onde empresa já está implantada

my first regression (in RM): load data

- data file: regression data.xlsx

- load data into repositor
  1. choose file
  2. choose worksheet
     - housing
  3. indicate row with column names (if any)
     - first row
  4. indicate column with ids of rows (if any)
     - none
  5. indicate column with target variable (if any)
     - last one
  6. give name
     - boston-housing
  7. finish
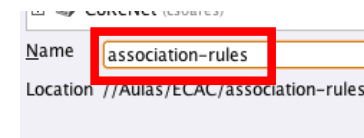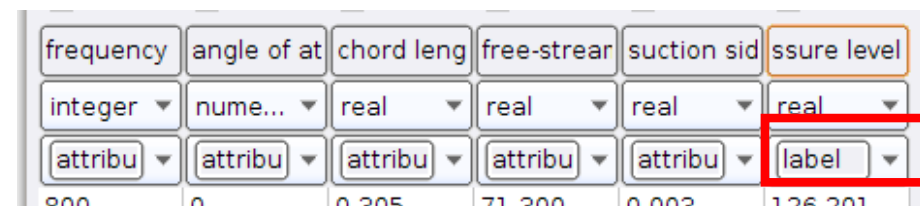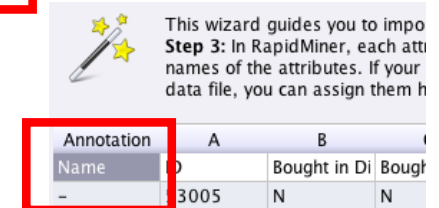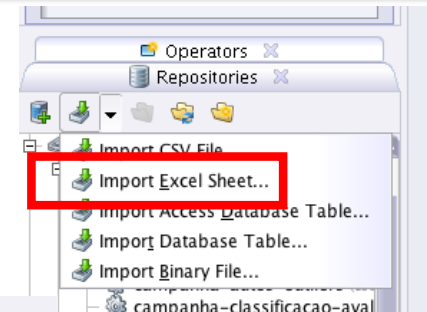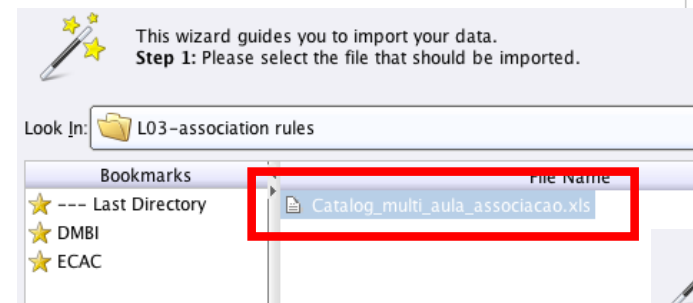
construir processo de regressão no rapid miner

- carregar dados do repositório
- aplicar algoritmo de regressão linear
  - ex. operators ➔ modeling ➔ classification and regression ➔ function fitting ➔ linear regression
- executar processo
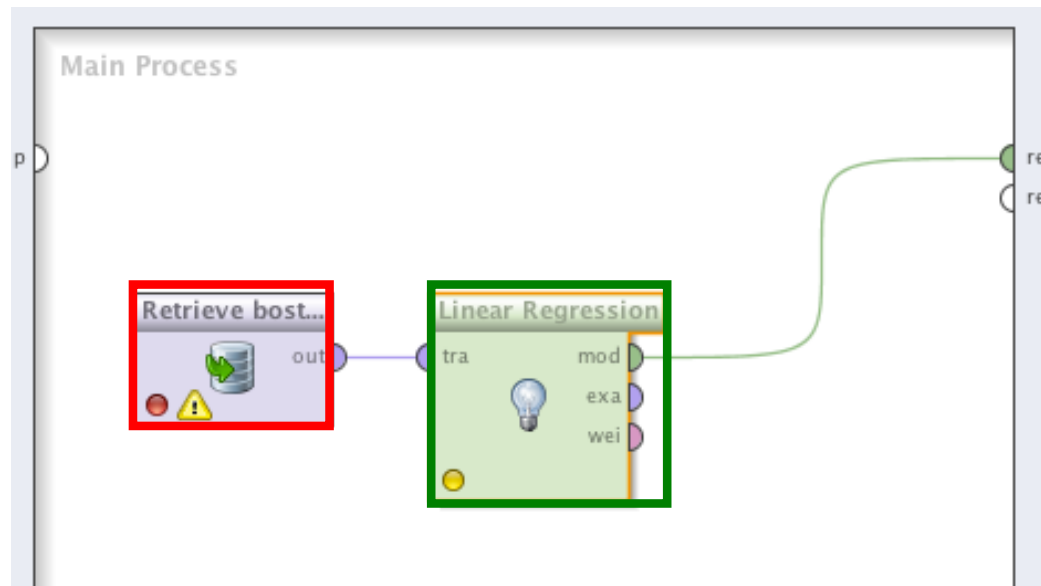
analyze linear regresssion model

- assumes variables are not correlated
  - influence of each variable explained separately
  - coefficients are not influenced by changes in the set of independent variables
- variation depends on magnitude of correlation
  - sign might change!

- ... but empirical results indicate robustness

○ Table View ◉ Text View ○ Annotations

LinearRegression

```
- 0.108 * CRIM
+ 0.045 * ZN
+ 0.018 * INDUS
+ 2.661 * CHAS
- 17.655 * NOX
+ 3.822 * RM
- 1.459 * DIS
+ 0.304 * RAD
- 0.012 * TAX
- 0.978 * PTRATIO
+ 0.009 * B
- 0.521 * LSTAT
+ 36.696
```

linear regression

- simple case: 2 variables
  x and *y*

- equation of the line

$$y = f(x)$$
$$= b_0 + b_1 x$$
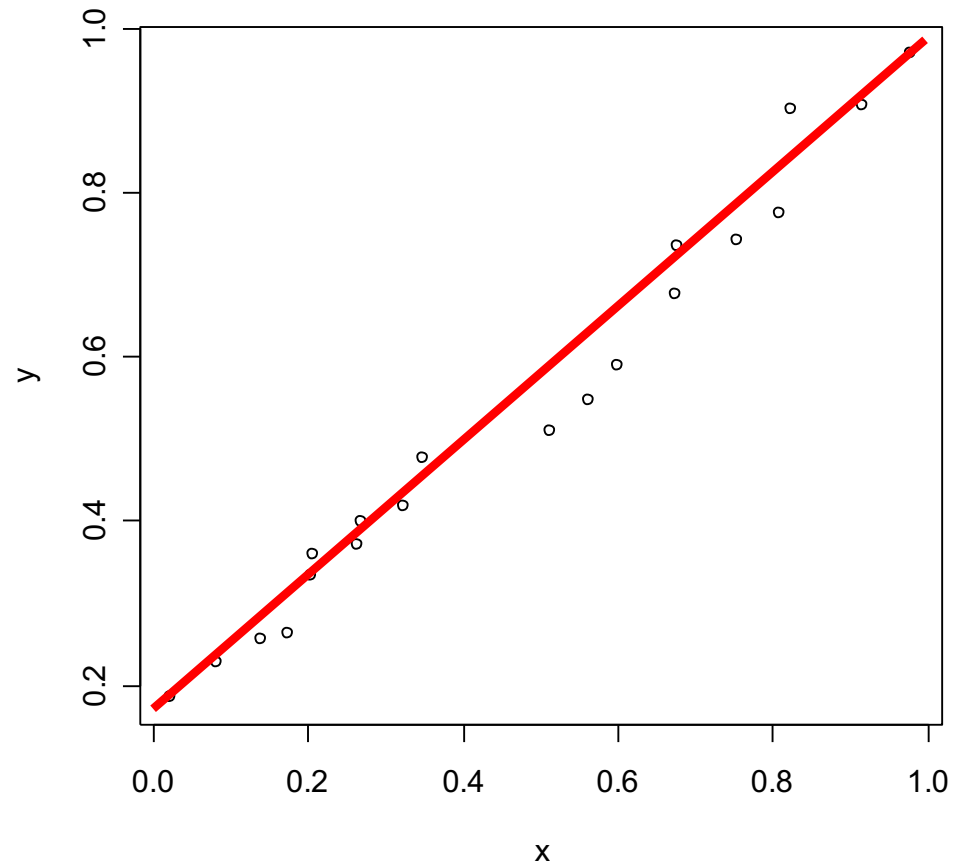
interpretion of coefficients

$$y = b_0 + b_1 x$$
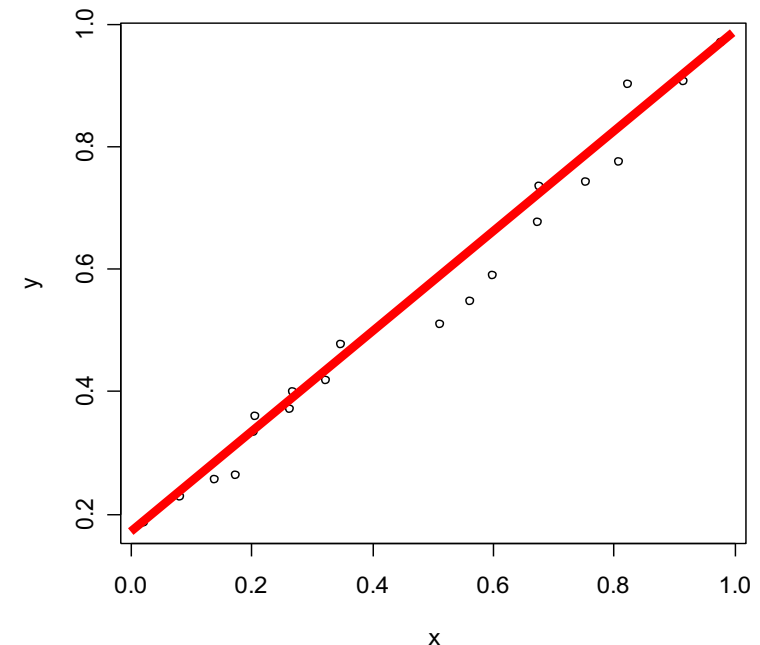
- *b_0*: intersection of the line with the *yy* axis
    - frequently hard to interpret
- *b_1*: slope of the line
    - variation of the value of *y* given an increase of 1 unit of *x*

make prediction for new examples

given a neighborhood with the following characteristics

CRIM	0,04294
ZN	28
INDUS	15,04
CHAS	0
NOX	0,464
RM	6,249
AGE	77,3
DIS	3,615
RAD	4
TAX	270
PTRATIO	18,2
B	396,9
LSTAT	10,59

… what is the predicted value?

○ Table View ⦿ Text View ○ Annotations

LinearRegression

```
−  0.108 * CRIM
+  0.045 * ZN
+  0.018 * INDUS
+  2.661 * CHAS
−  17.655 * NOX
+  3.822 * RM
−  1.459 * DIS
+  0.304 * RAD
−  0.012 * TAX
−  0.978 * PTRATIO
+  0.009 * B
−  0.521 * LSTAT
+  36.696
```

Simple linear regression: estimating parameters

$$y = b_0 + b_1 x$$

$$\widehat{b_1} = \frac{S_{XY}}{S_{XX}}$$

where $\widehat{b_1}$ is an estimate of b

$$S_{XY} = \sum_{i=1}^{n} [(X_i - \overline{X}).(Y_i - \overline{Y})]$$

$$S_{XX} = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

- $\widehat{b_1}$ should be statistically significantly different from zero
  - if not, there is no meaningful dependency between *Y* and X
  - this should be tested

$$\widehat{b_0} = \overline{Y} - \hat{\beta}.\overline{X}$$

where $\widehat{b_0}$ is an estimate of b_0

- $\widehat{b_0}$ may or may not be statistically significantly different from zero
  - If not there is no evidence that *Y≠0* when *X=0.*
  - … which could make sense
    - e.g. value of a customer with 0 income
  - … or not…
    - e.g. minimum sales of a product without shelf space

Simple linear regression: assumptions

- Linear relationship between *x* and *y*
  - also additive
- Errors
    - i.e. unexplained variation in *y*
  - ... are independently and identically distributed
  - ... homoscedasticity
    - constant variance
  - ... normally distributed

gravar modelo com rapid miner

- gravar modelo no repositório
  - operador store
    - indicar o nome a dar ao objeto a criar como parâmetro
  - executar

carlos soares / MEST

predict new examples with rapid miner

- **get new data**
  - import
    - same spreadsheet
    - "housing (new)" worksheet
    - target value is not filled but it's role is still "label"
- **get model**
  - previously stored
- **apply model to new data**
  - operator "Apply Model"
- **obtain predictions**
  - run process

Row No.	MEDV	prediction(...	CRI
1	?	25.438	2.92
2	?	37.265	2.01
3	?	23.163	1.80
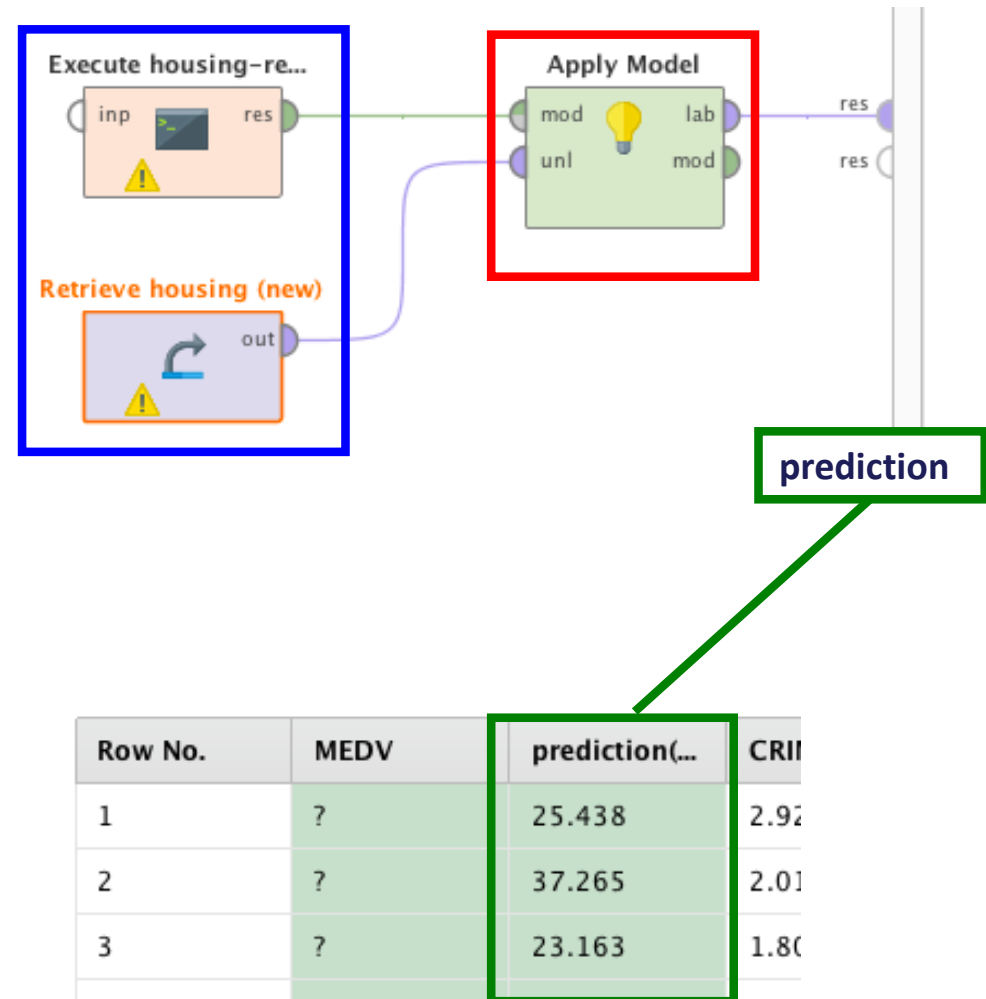
AVALIAÇÃO DE MODELOS DE REGRESSÃO

regressão: resumo (até agora)

exemplos com valor conhecido da variável-objetivo

Comprou	Idade	Rendimento	Ag.fam	Vendas anteriores	Última Venda
não	37	49000	2	1	42000
sim	43	68000	3	0	0
sim	42	61000	4	0	0
sim	26	52000	2	0	0
sim	40	64000	1	1	21000
sim	38	52000	1	0	0
sim	45	43000	4	1	47000
sim	35	45000	2	1	34000
não	39	43000	2	0	0

(novos) exemplos com valor desconhecido da variável-objetivo

	A	B	C	D	
1	Comprou	Idade	Rendimento	Ag.fam	Vendas
2		41	50000	2	
3		39	68000	2	
4		58	61000	4	
5		26	25000	3	
6		21	50000	1	
7		38	43000	2	
8		44	43000	4	
9		27	47000	2	
10		70	23000	2	

1

2
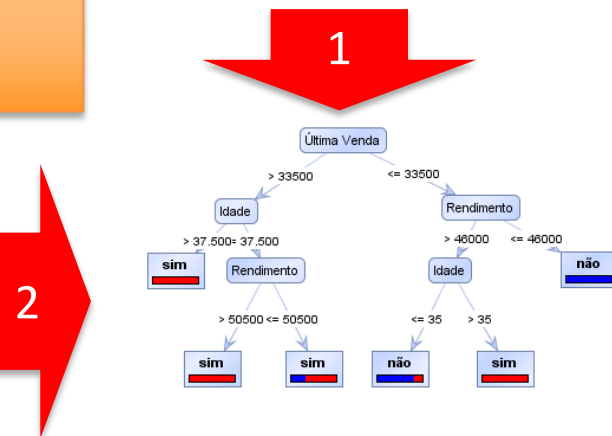
previsões para os (novos) exemplos com valor desconhecido da variável-objetivo

row no.	Comprou	prediction(...	confidence(...	confidence(...	Ida
1	?	sim	0	1	41
2	?	sim	0	1	39
3	?	sim	0	1	58
4	?	não	1	0	26
5	?	não	0.818	0.182	21
6	?	não	1	0	38
7	?	sim	0	1	44
8	?	não	0.818	0.182	27
9	?	não	1	0	70

3

usariam as previsões feitas por este modelo?

gps

- regression
- linear regression
- evaluation of regressors
  - measures
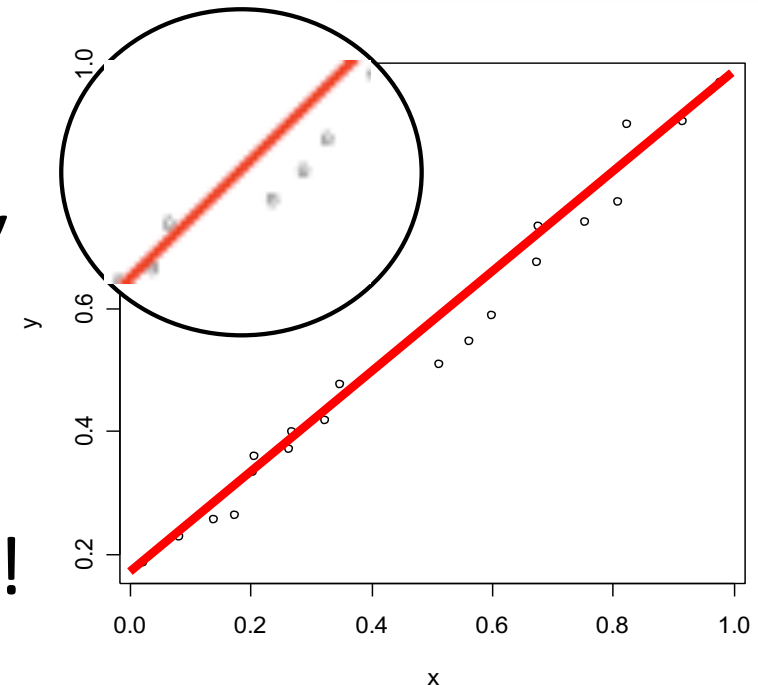 - methodology

prediction and evaluation

- given the value of *x*

- model estimates the value of *y*

$$\hat{y} = b_0 + b_1 x$$

- but the estimate is not perfect!

$$\hat{y} - y$$

- error:
  - y : true value
  - ŷ : value estimated by the model

analysis of evaluation measures

- average error
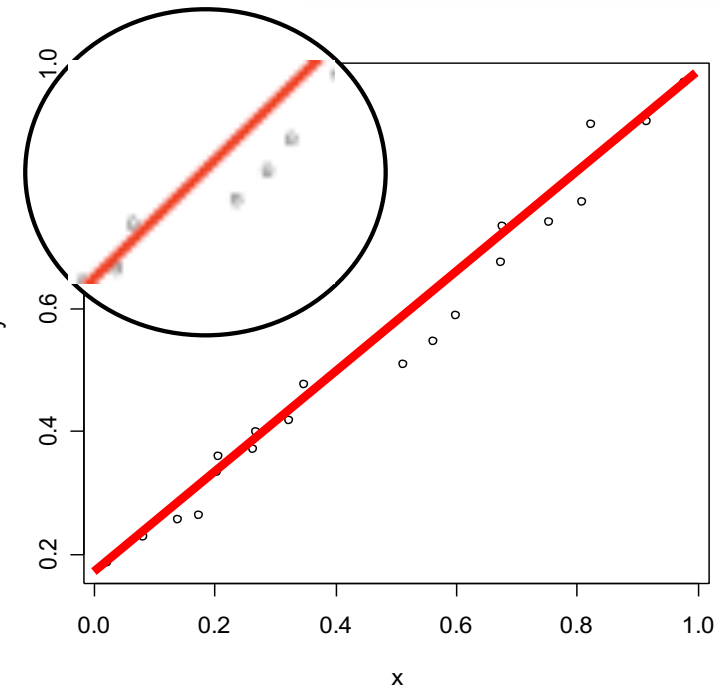 - do not use!

$$\frac{1}{m} \sum_i \hat{y}_i - y_i$$

- mean absolute deviation
  - estimates "typical" error

$$\frac{1}{m} \sum_i |\hat{y}_i - y_i|$$

- mean squared error
  - gives more importance to larger errors
  - ... possible dominated by a few errors

$$\frac{1}{m} \sum_i (\hat{y}_i - y_i)^2$$

- values depend on the scale of the target variable
  - good or bad?
    - business perspective
    - but is it really the relationship between *x* and *y*?

baseline

- if nothing is known about the new case

- ... what is the best prediction we can make?
  - random vs average

$$\hat{y}_i = \overline{y}$$

- baseline

- regression is only useful if the error is less than the baseline error
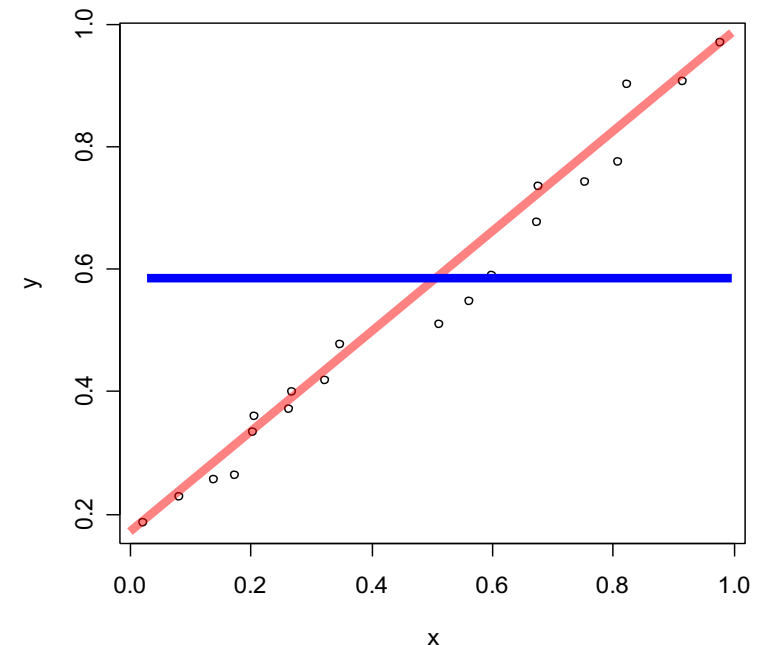 - eg MSE

$$\frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\overline{y} - y_i)^2}$$

0 if prediction model is perfect

]0,1[if it is useful
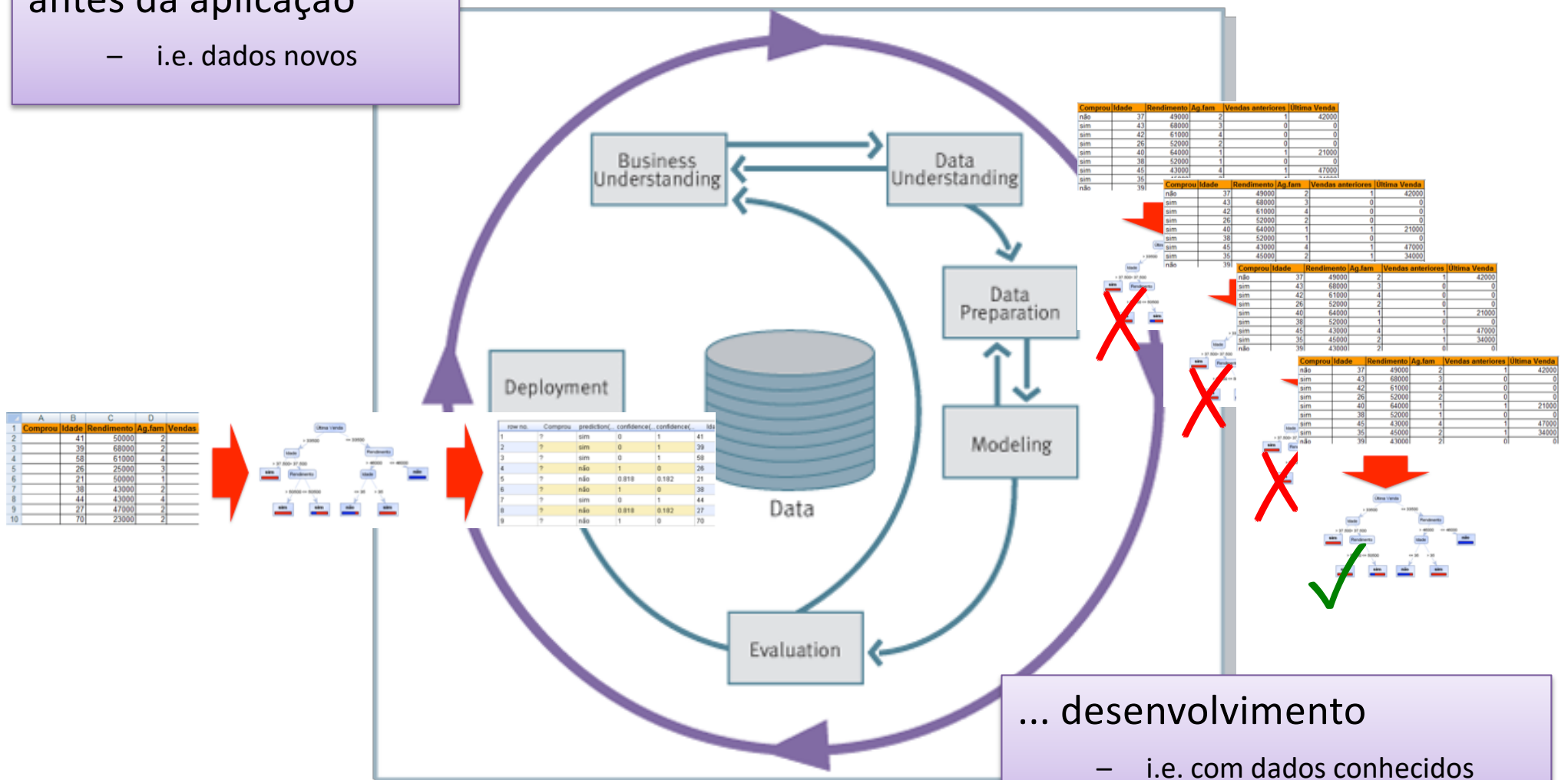
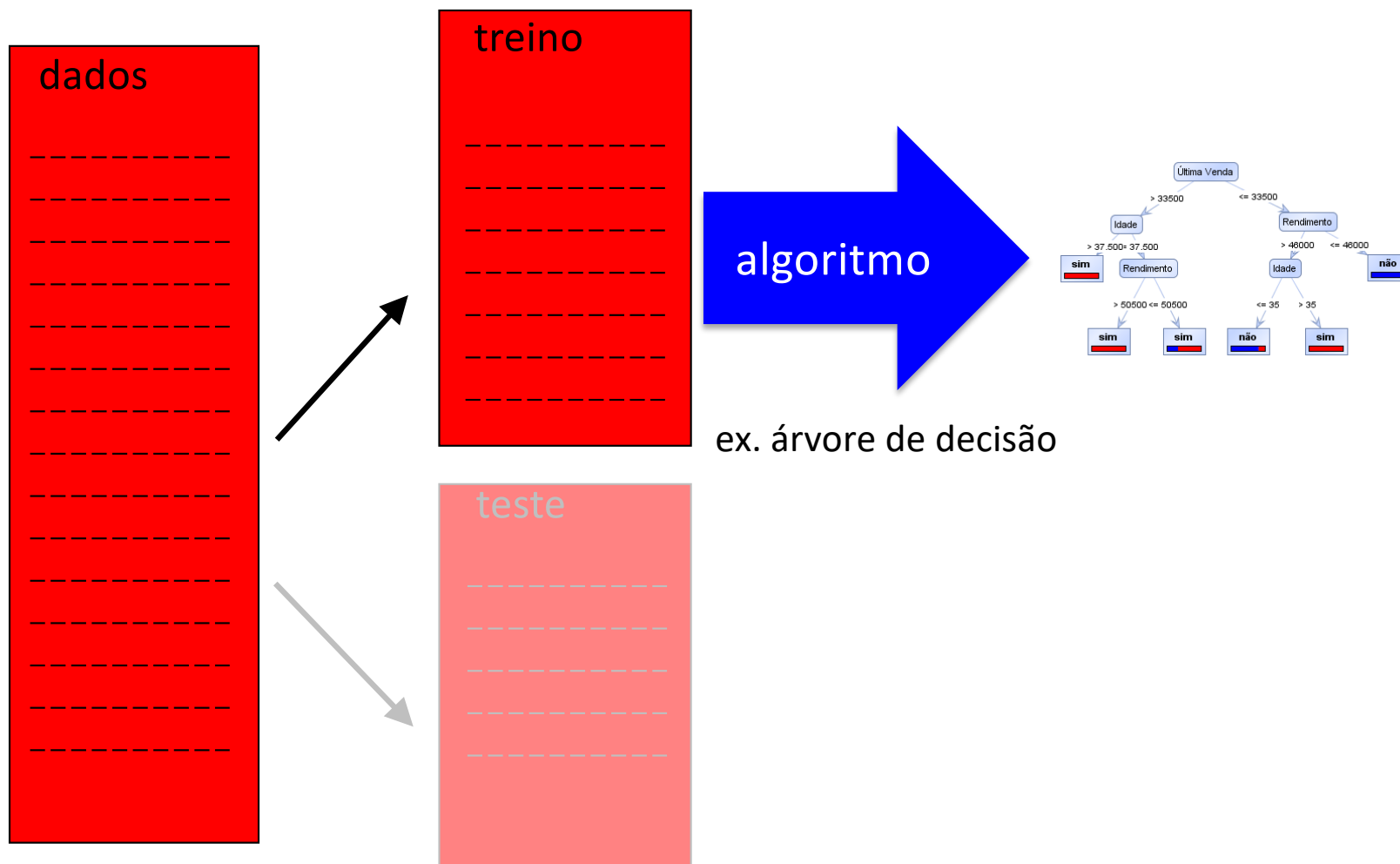1 if it is the same as the baseline

>1 if it is worse than the baseline

desenvolvimento de modelos
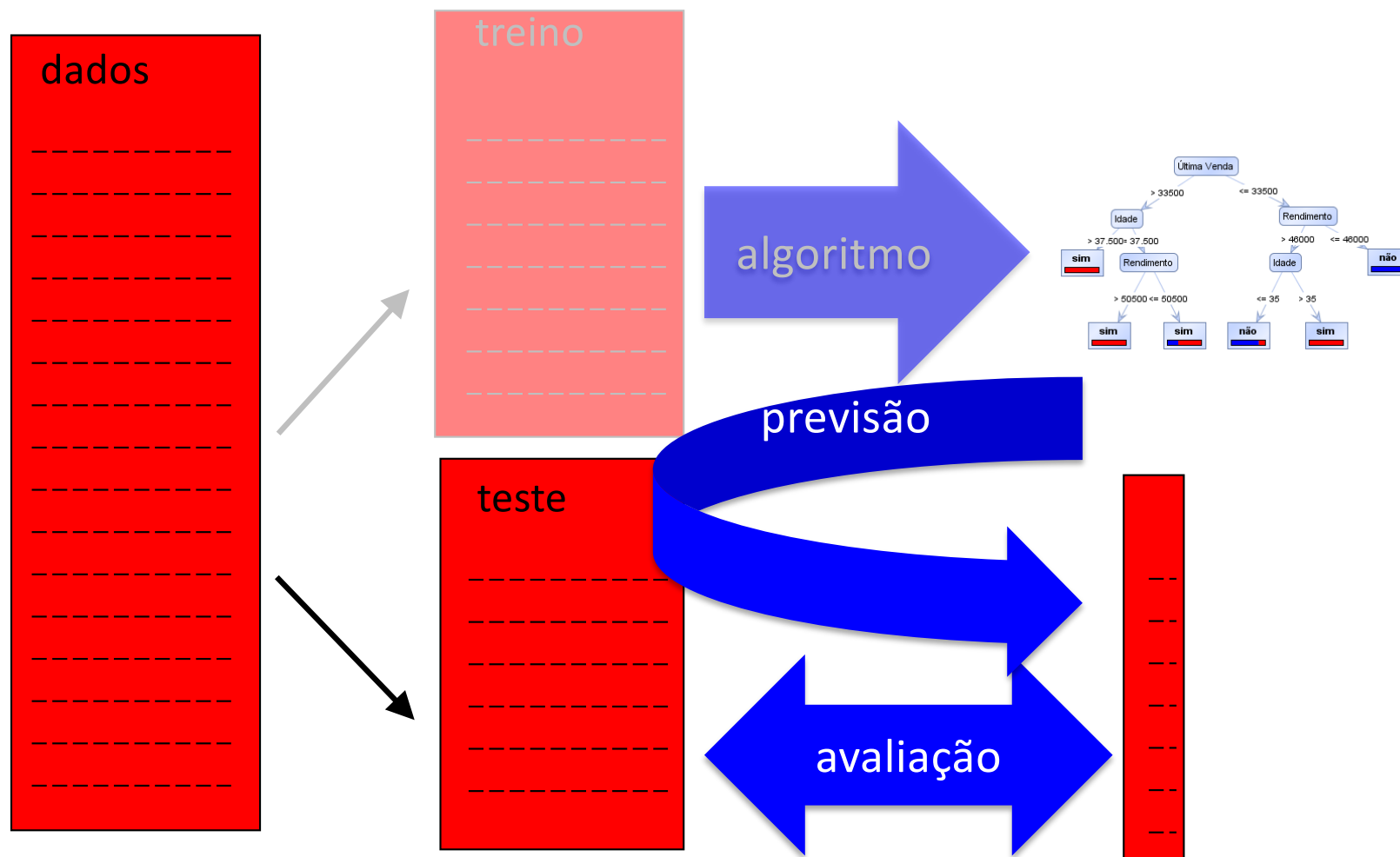

antes da aplicação
- i.e. dados novos

... desenvolvimento
- i.e. com dados conhecidos

metodologia de avaliação: treino

dados

treino

algoritmo

ex. árvore de decisão

teste

metodologia de avaliação: teste

ex. taxa de acerto

avaliar modelo de regressão em rapid miner (1/3)

- operador split validation
- sub-processo
  - operador que contém outros operadores
  - duplo clique para entrar

- distribuir aleatoriamente os dados por conjunto de treino e conjunto de teste
  - proporção
    - 70% dos casos para treino
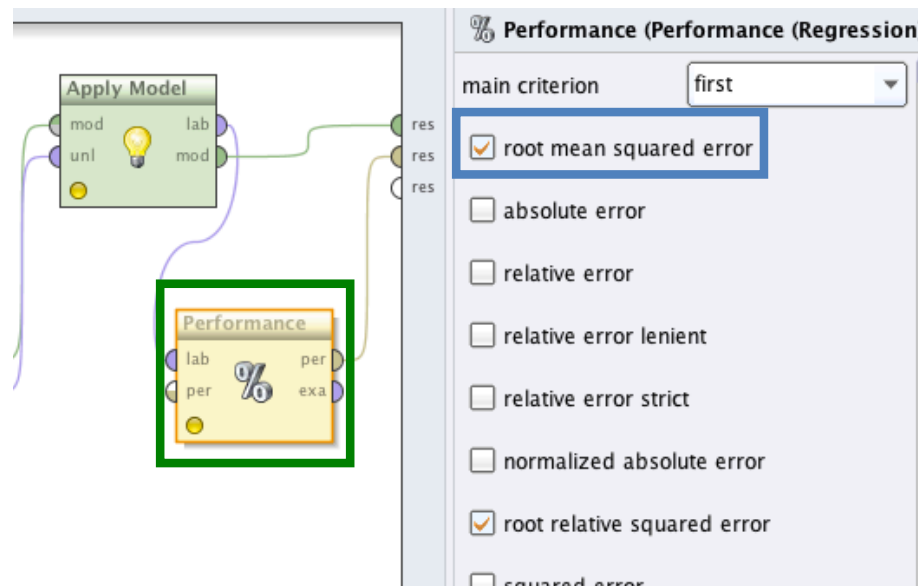   - 30% dos casos para teste

carlos soares / MEST

25

estimate predictive error with rapid miner (2/3)

- split validation
  - different operations applied to training and test data
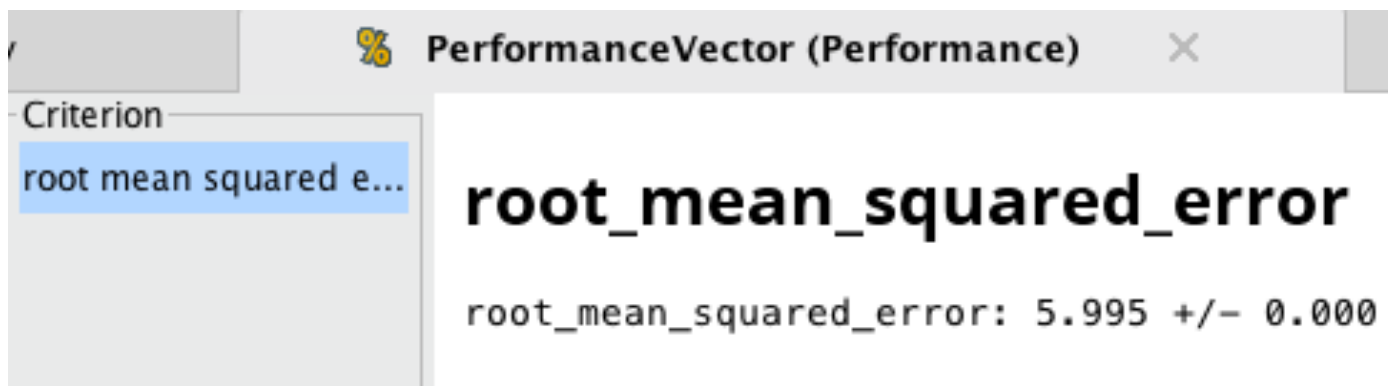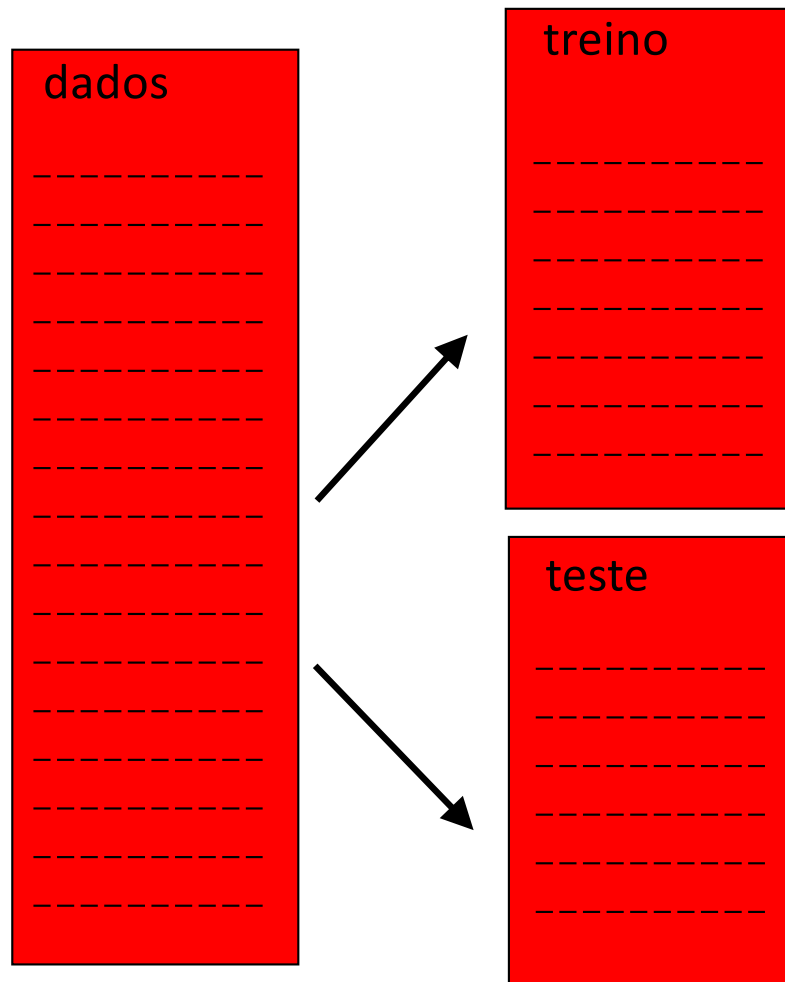 - model obtained with train data is applied to test data
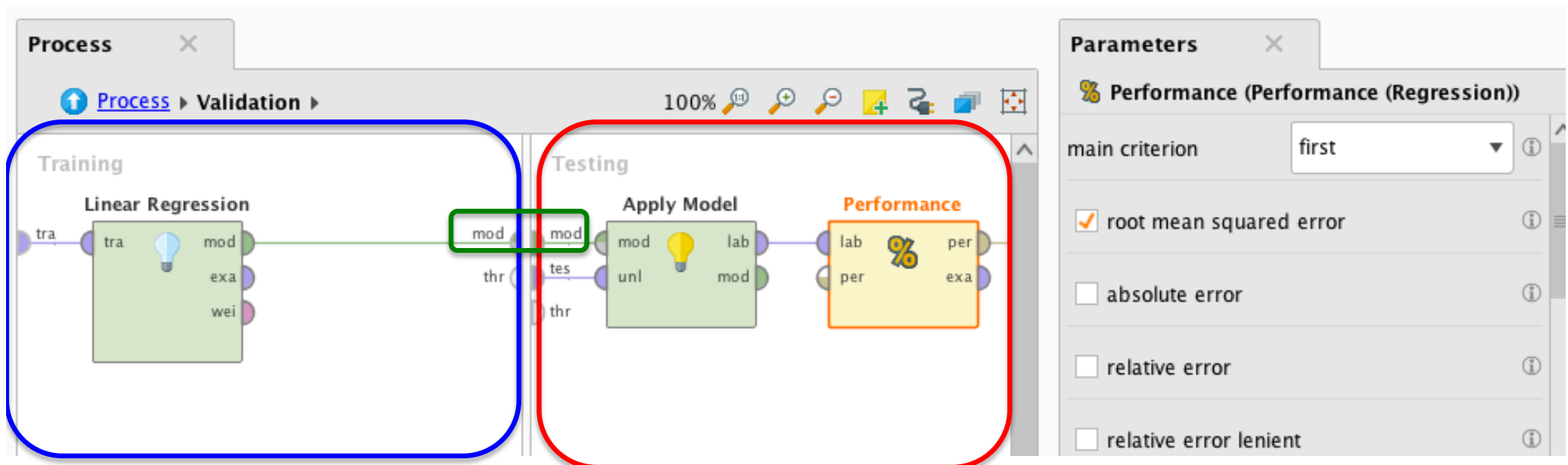
avaliar modelo de regressão em rapid miner (3/3)

- operador "Performance (Regression)"
- escolher medidas RMSQ

estimativa de desempenho do modelo de regressão

metodologia de avaliação: treino

- treino
  - para obter o modelo automaticamente

- teste
  - estimar o valor do modelo em novos casos
  - assume que os novos dados terão uma distribuição idêntica aos de treino
    - não funciona se houver alterações na distribuição: ex: inflação

short detour:
brief introduction to bias
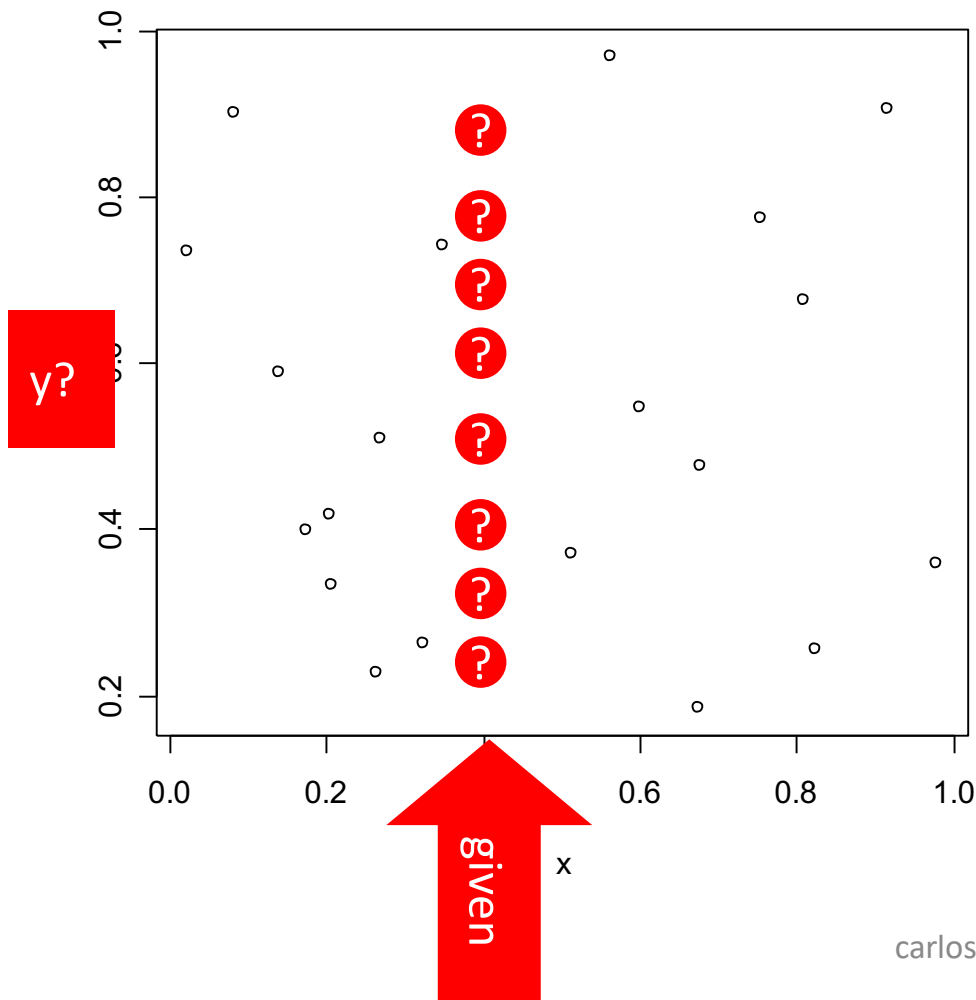
- y is customer value
- x is family income
  - and other characteristics
  - ... only 1 here for simplicity

- why is it harder to predict y now?