Data-Science 1

machine learning

Karel de Grote Hogeschool

24/04/2024

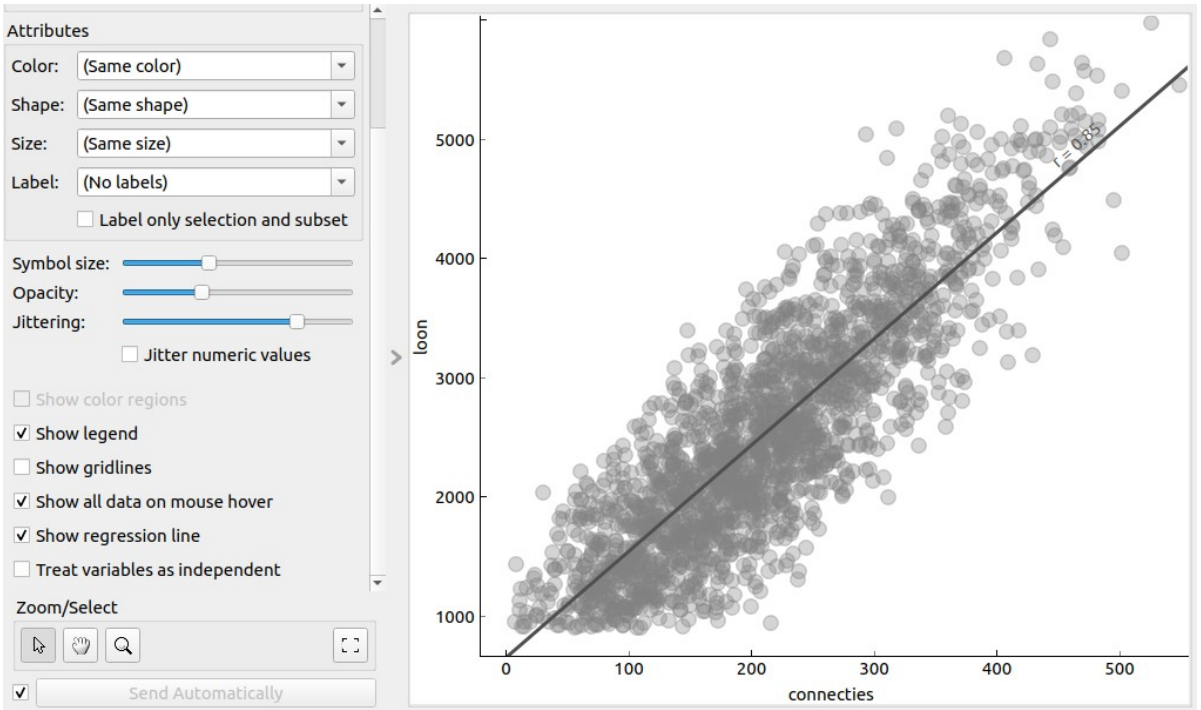
Inhoud

- regressie en metrieken
  - multivariate regressie
  - evaluatiemetrieken
    - MSE
    - RMSE
    - MAE
    - MAPE

Multivariate Lineaire Regressie

Lineaire regressie

- wat was (bivariate) lineaire regressie ook al weer?



Lineaire regressie

- lineaire regressie is een (eenvoudige) machine learning techniek:
  - je geeft voorbeelden uit het verleden (puntenwolk)
  - je stelt een model op om voorspellingen te kunnen doen (een rechte lijn met slope en intercept)
  - je gebruikt het model om voorspellingen te doen

Multivariate lineaire regressie

- je kan lineaire regressie ook toepassen op meerdere variabelen
- voorbeeld:

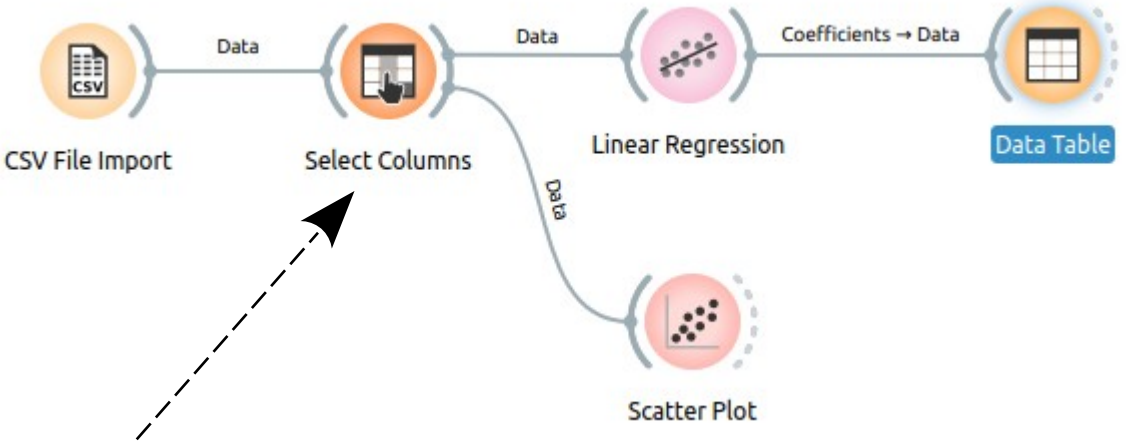
	verkoopresultaat	leeftijd	inkomen	werkervaring	klantencontacten
1	239.231	34	35125.3	3	21
2	244.051	33	46337.8	5	11
3	219.486	33	32189.6	6	9
4	309.835	32	55517	7	17
5	220.038	35	33362.8	5	12
6	212.289	28	46780.3	4	8
7	263.155	37	37993	4	19
8	271.657	45	37197.9	8	10
9	337.086	30	50566.6	9	20
10	312.401	37	41956.8	11	19
11	245.454	40	40179.9	3	17
12	217.946	33	46511.7	3	9
13	264.251	29	35866	6	14

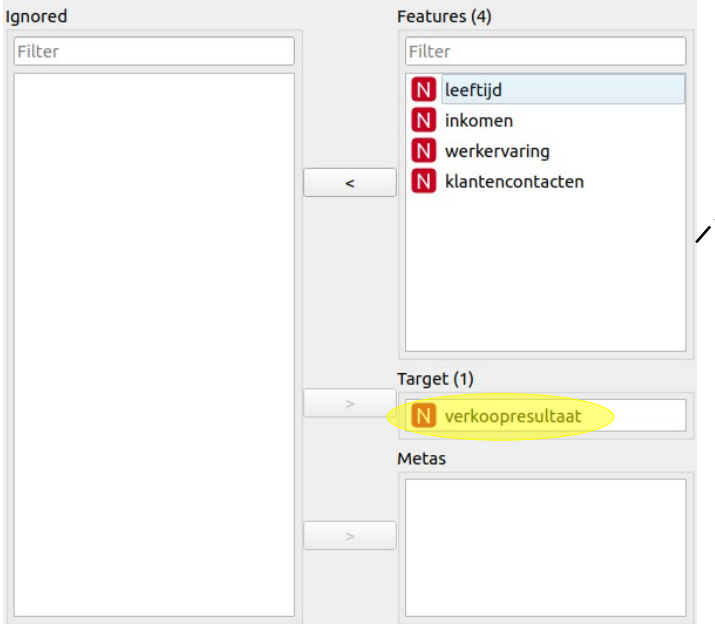
- verkoopresultaat = a + b * leeftijd + c * inkomen + d * werkervaring + e * klantencontacten

Multivariate lineaire regressie

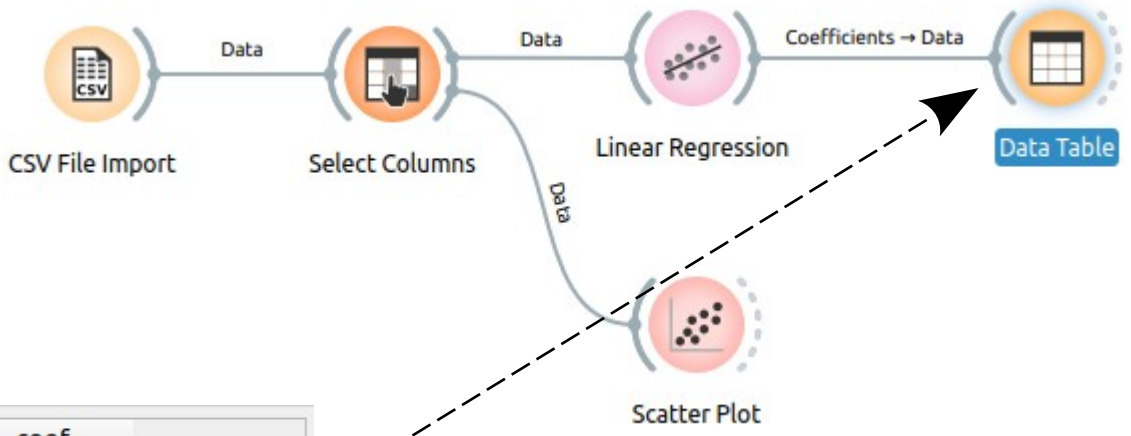
- verkoopresultaat = a + b * leeftijd + c * inkomen + d * werkervaring + e * klantencontacten
- dit heet een “lineaire combinatie” van de kolommen leeftijd, inkomen, werkervaring en klantencontacten
- opmerkingen
  - kolommen moeten minstens interval meetniveau hebben
  - iedere rij is nu ook een punt, maar in hoger-dimensionale ruimte...
  - we zoeken nu een “hypervlak” dat door de punten gaat

Orange


Data
Data
Coefficients → Data
CSV File Import
Select Columns
Linear Regression
Data Table
Data
Scatter Plot


Ignored
Filter
Features (4)
Filter
leeftijd
inkomen
werkervaring
klantencontacten
<
Target (1)
verkoopresultaat
Metas

Orange



Info

5 instances (no missing data)
1 feature
No target variable.
1 meta attribute

Variables

- [x] Show variable labels (if present)
- [ ] Visualize numeric values
- [x] Color by instance classes

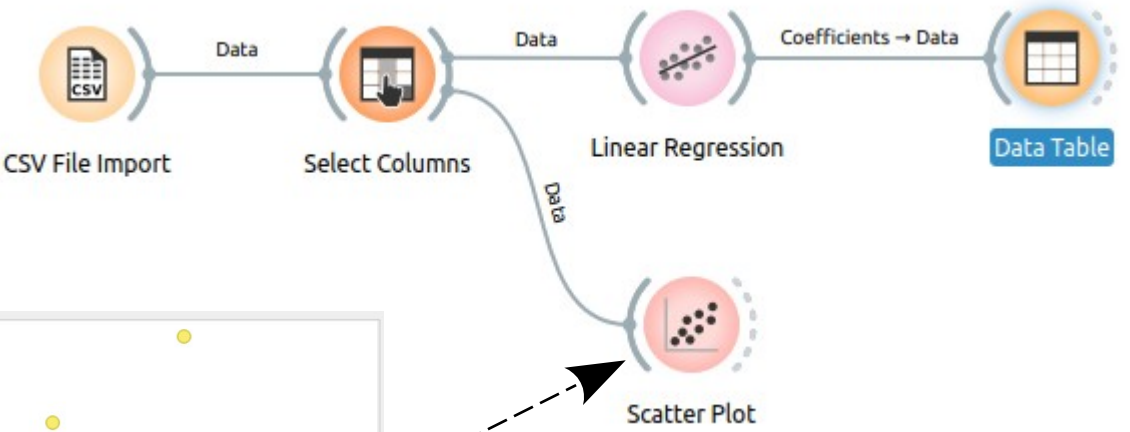
Selection

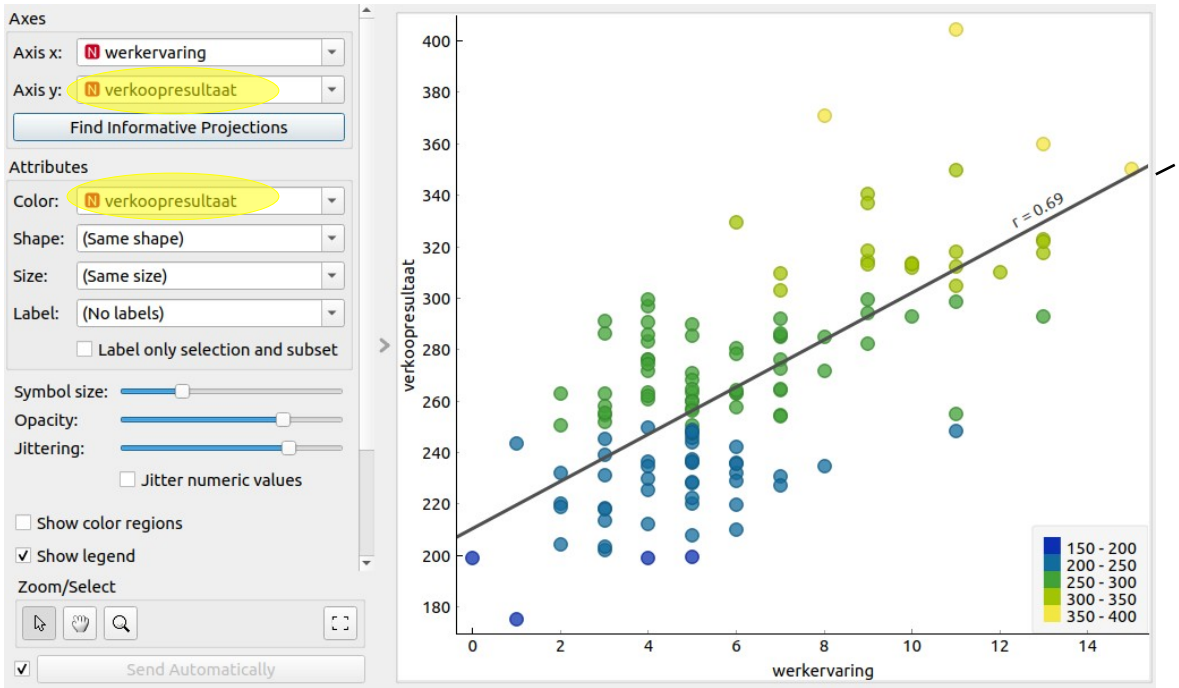
Restore Original Order

[x] Send Automatically

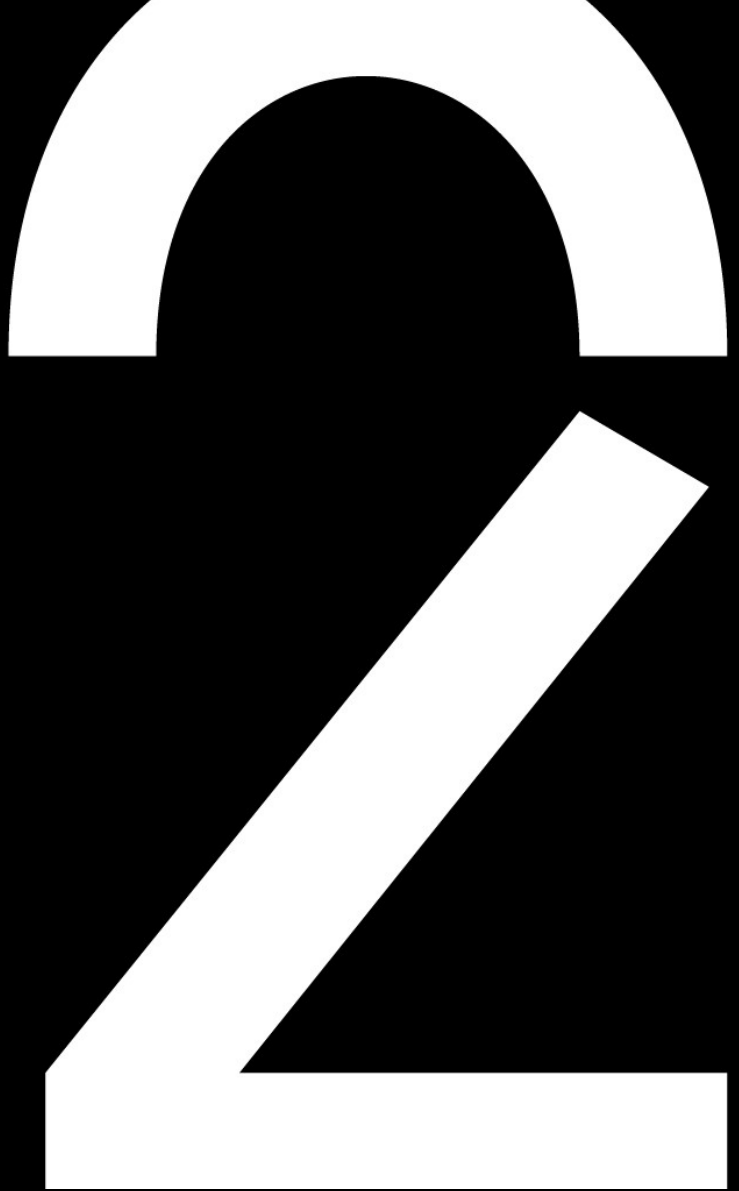
	name	coef
1	intercept	11.3227
2	leeftijd	1.27979
3	inkomen	0.00230116
4	werkervaring	9.22117
5	klantencon…	3.83109

Orange


Data
Data
Coefficients → Data
CSV File Import
Select Columns
Linear Regression
Data Table
Data
Scatter Plot


Axes
Axis x: werkervaring
Axis y: verkoopresultaat
Find Informative Projections
Attributes
Color: verkoopresultaat
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Show color regions
Show legend
Zoom/Select
Send Automatically
verkoopresultaat
werkervaring
r = 0.69
150 - 200
200 - 250
250 - 300
300 - 350
350 - 400

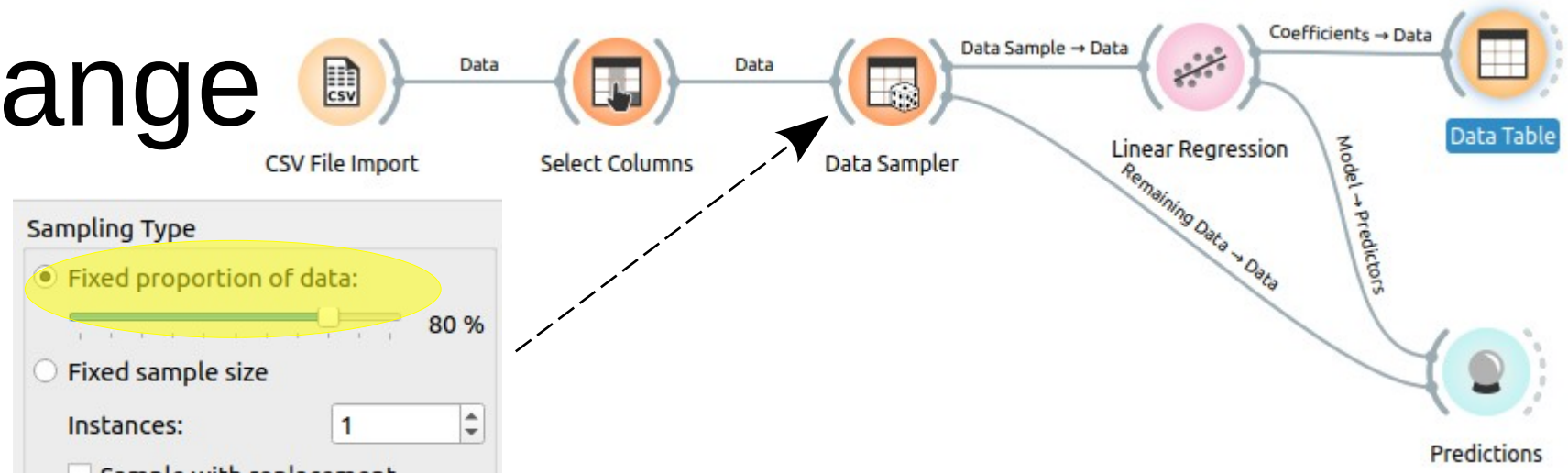
Evaluatiemetrieken

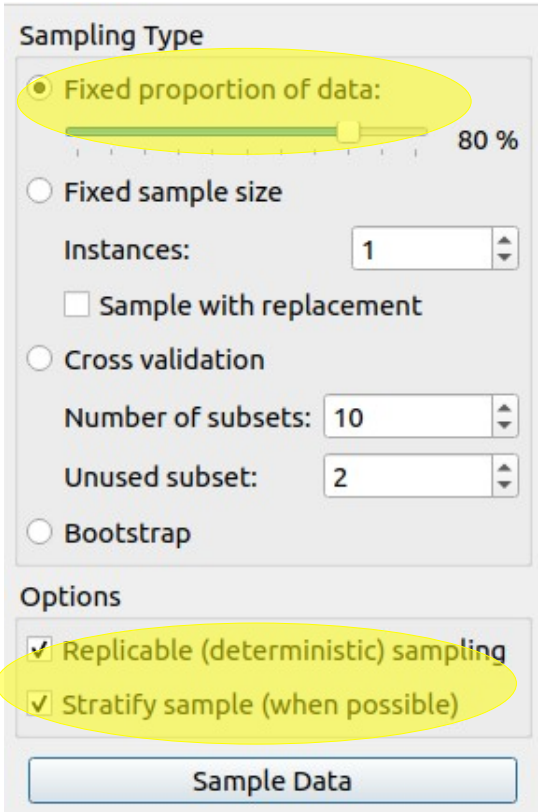


Multivariate lineaire regressie

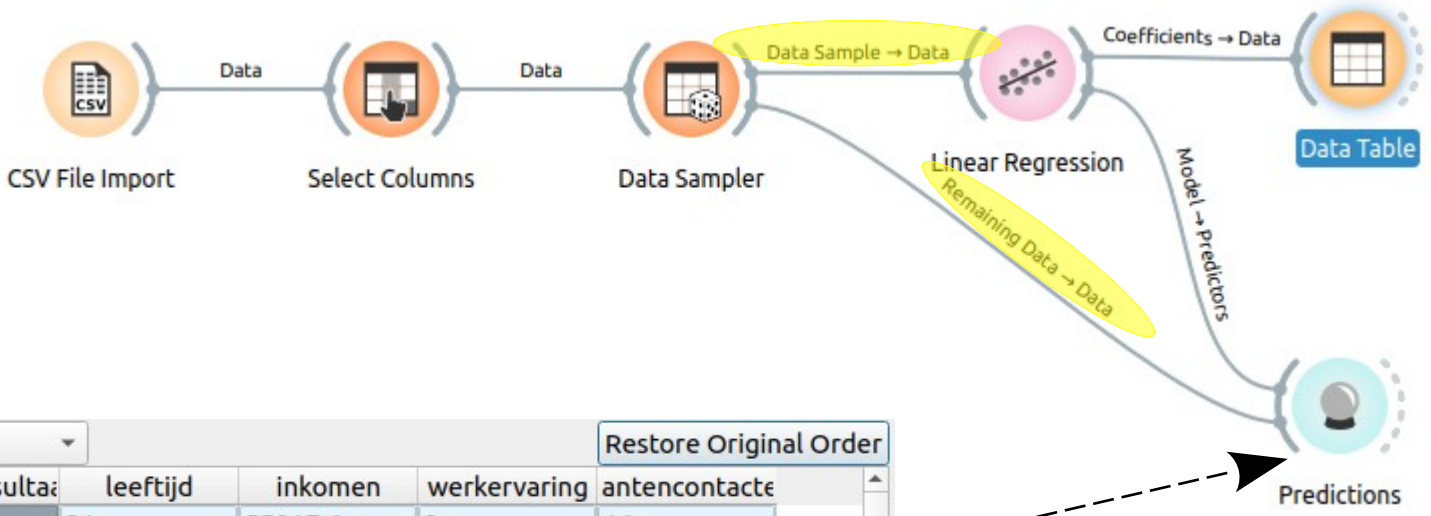
- hoe weten we hoe goed dit model is?
  - waarde van RMSE geeft enkel weer hoe goed het model presteert op de gegeven data
  - er kan dus overfitting ontstaan wanneer de RMSE heel klein wordt
  - hoe weten we of het ook goed is voor nieuwe data?
  - oplossing: splits de dataset op in 2 delen
    - training dataset
    - test dataset

Orange


CSV
Data
Data
Data Sample → Data
Coefficients → Data
CSV File Import
Select Columns
Data Sampler
Linear Regression
Data Table
Remaining Data → Data
Model → Predictors
Predictions


Sampling Type
Fixed proportion of data:
80 %
Fixed sample size
Instances:
1
Sample with replacement
Cross validation
Number of subsets:
10
Unused subset:
2
Bootstrap
Options
Replicable (deterministic) sampling
Stratify sample (when possible)
Sample Data

Orange



Shown regression error: Difference

Restore Original Order

	Linear Regression	error	erkoopresultaa	leeftijd	inkomen	werkervaring	antencontacte
1	259.493	3.9665	255.526	31	55917.6	3	14
2	321.907	8.51234	313.394	40	45700.5	10	16
3	272.18	23.808	248.372	33	33910.5	11	10
4	222.278	-13.674	235.952	29	33513.7	6	11
5	247.496	3.44584	244.051	33	46337.8	5	11

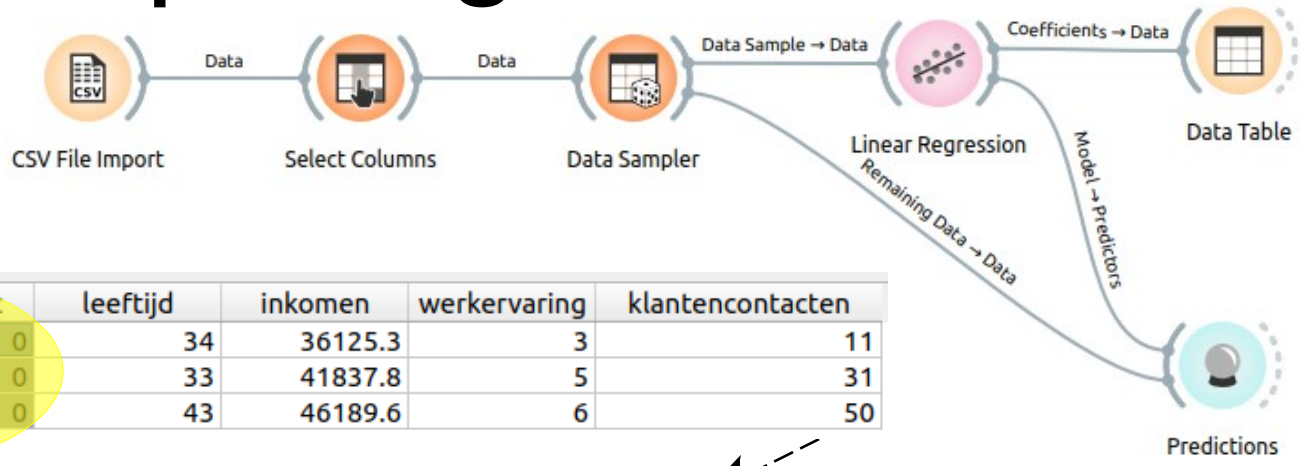
☑ Show perfomance scores

Model	MSE	RMSE	MAE	MAPE	R2
Linear Regression	205.085	14.321	12.289	0.048	0.869

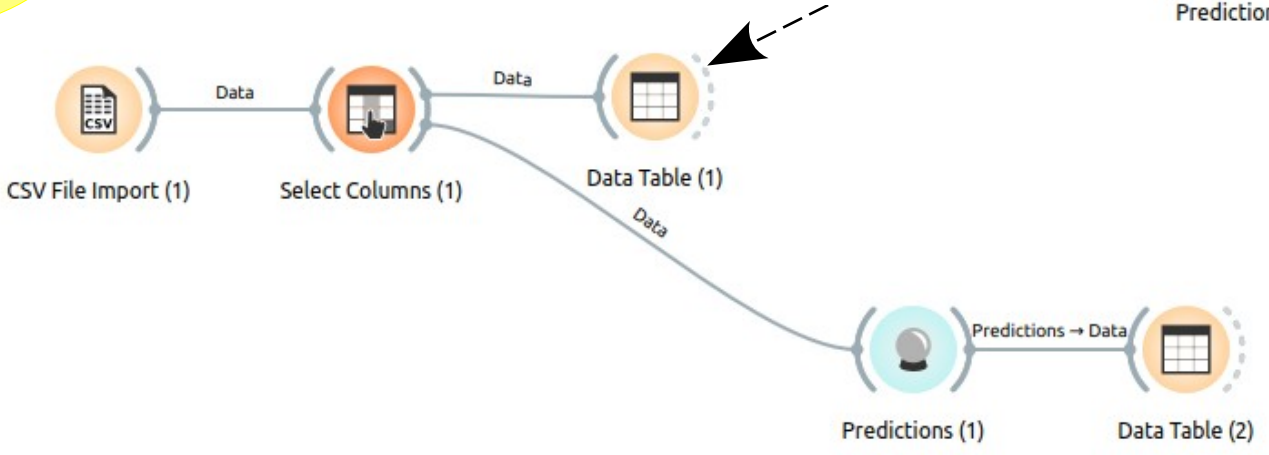
Metrieken

- Mean Squared Error: $MSE = \frac{1}{n} \sum e_i^2$
- Root Mean Squared Error: $RMSE = \sqrt{\frac{1}{n} \sum e_i^2}$
- Mean Absolute Error: $MAE = \frac{1}{n} \sum |e_i|$
- Mean Absolute Percentage Error: $MAPE = \frac{1}{n} \sum \left| \frac{e_i}{x_i} \right|$

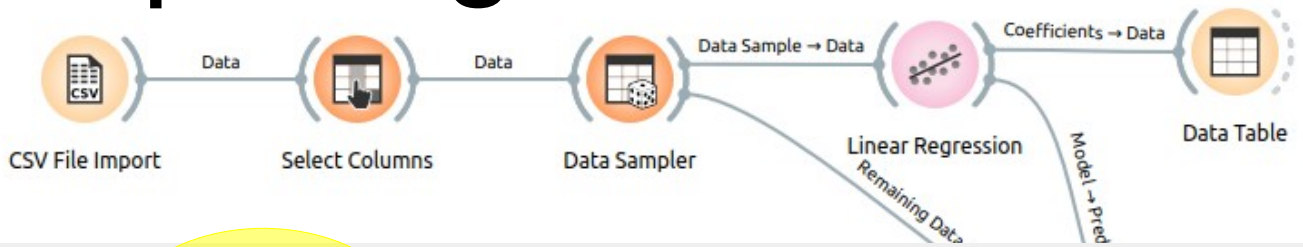
Voorspellingen maken



	verkoopresultaat	leeftijd	inkomen	werkervaring	klantencontacten
1	0	34	36125.3	3	11
2	0	33	41837.8	5	31
3	0	43	46189.6	6	50

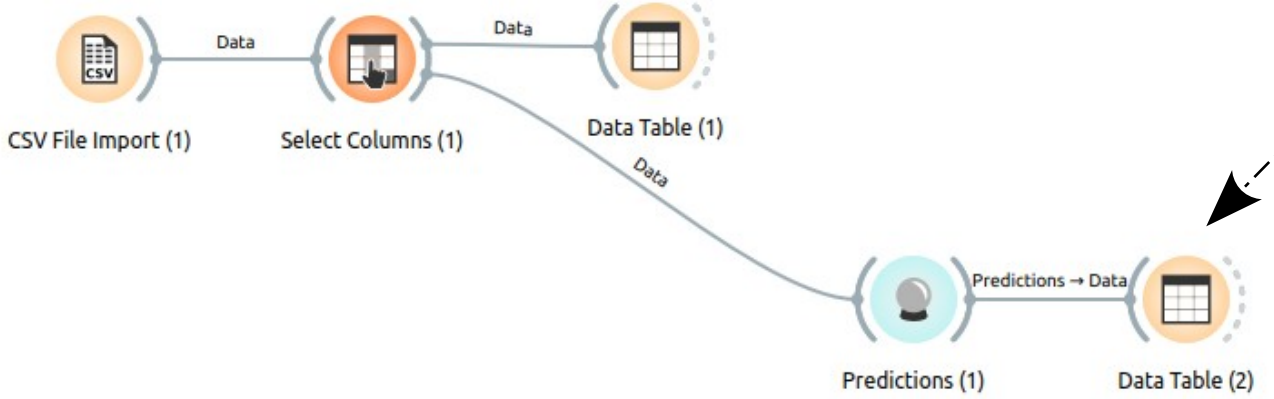


Voorspellingen maken



	verkoopresultaat	Linear Regression	Linear Regression (error)	leeftijd	inkomen	werkervaring	klantencontacten
1	0	206.127	206.127	34	36125.3	3	11
2	0	313.906	313.906	33	41837.8	5	31
3	0	419.308	419.308	43	46189.6	6	50

Predictions



Oefeningen

Oefeningen

- Zie Canvas
  - Revenue