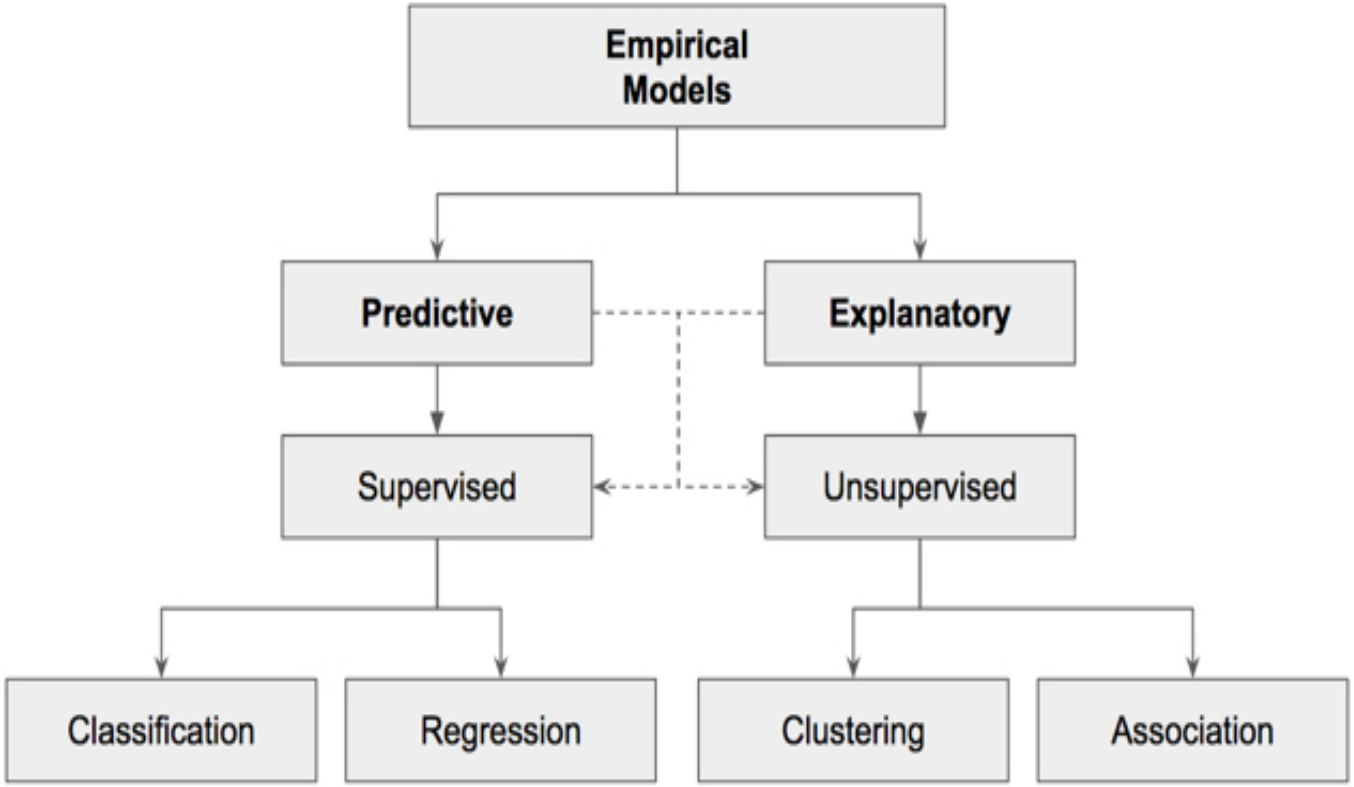
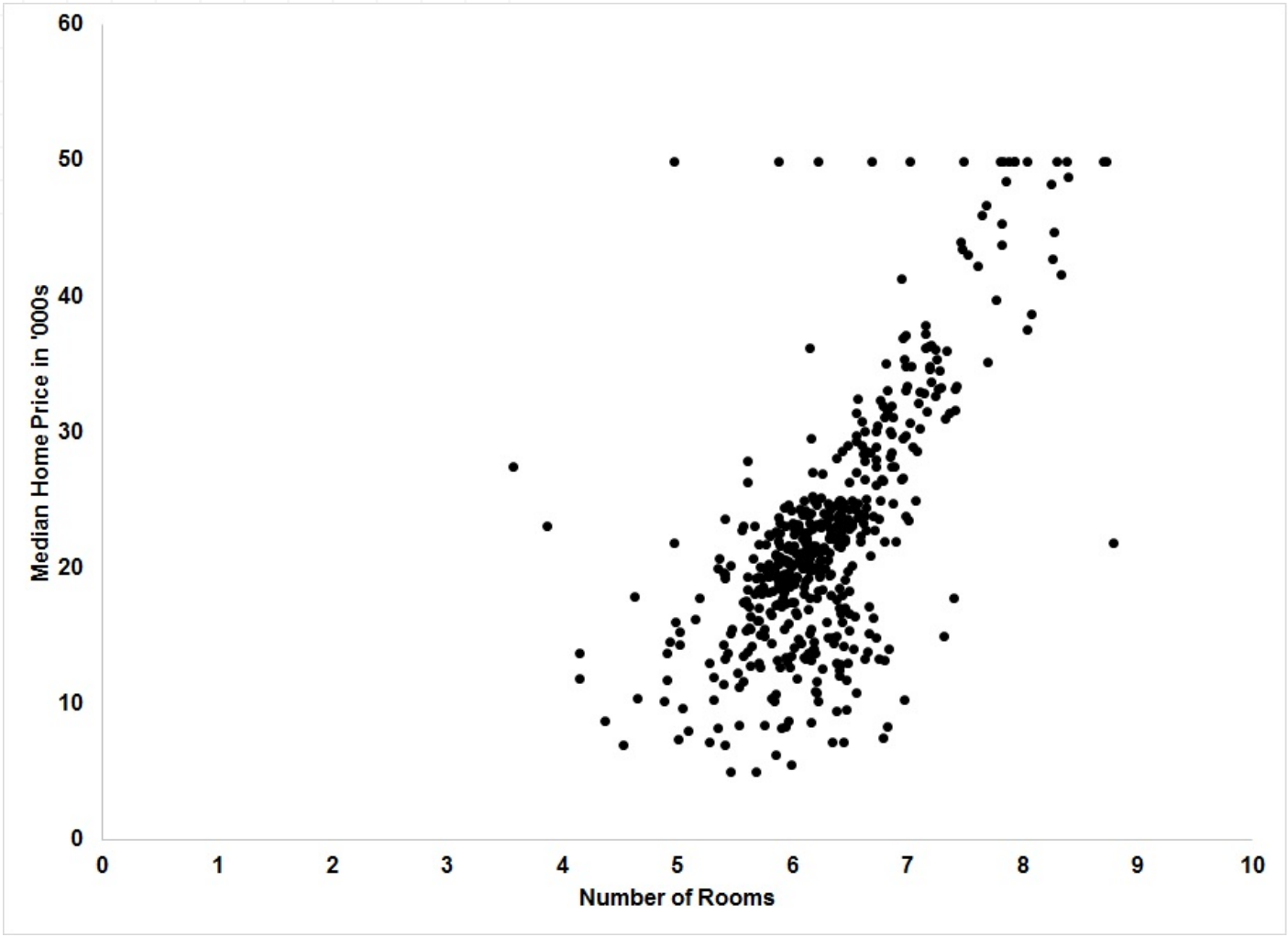
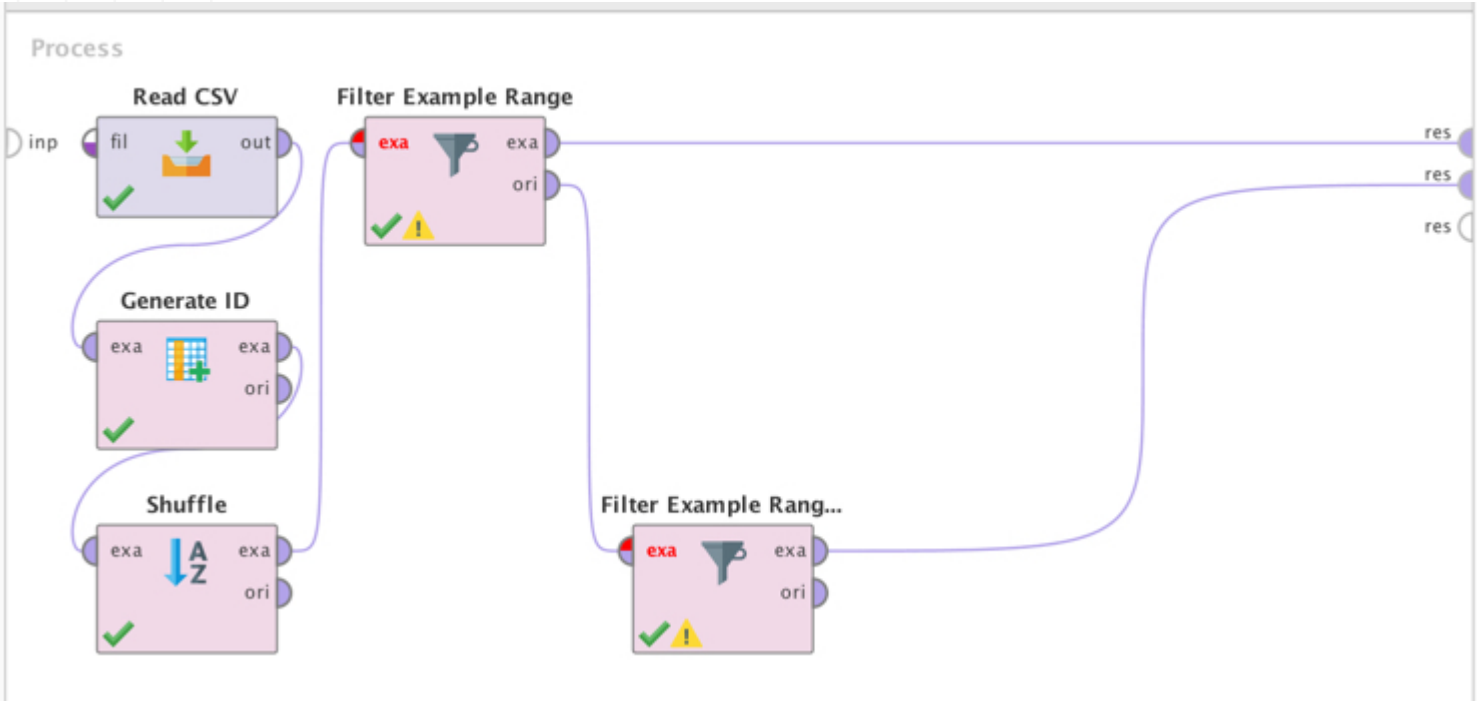
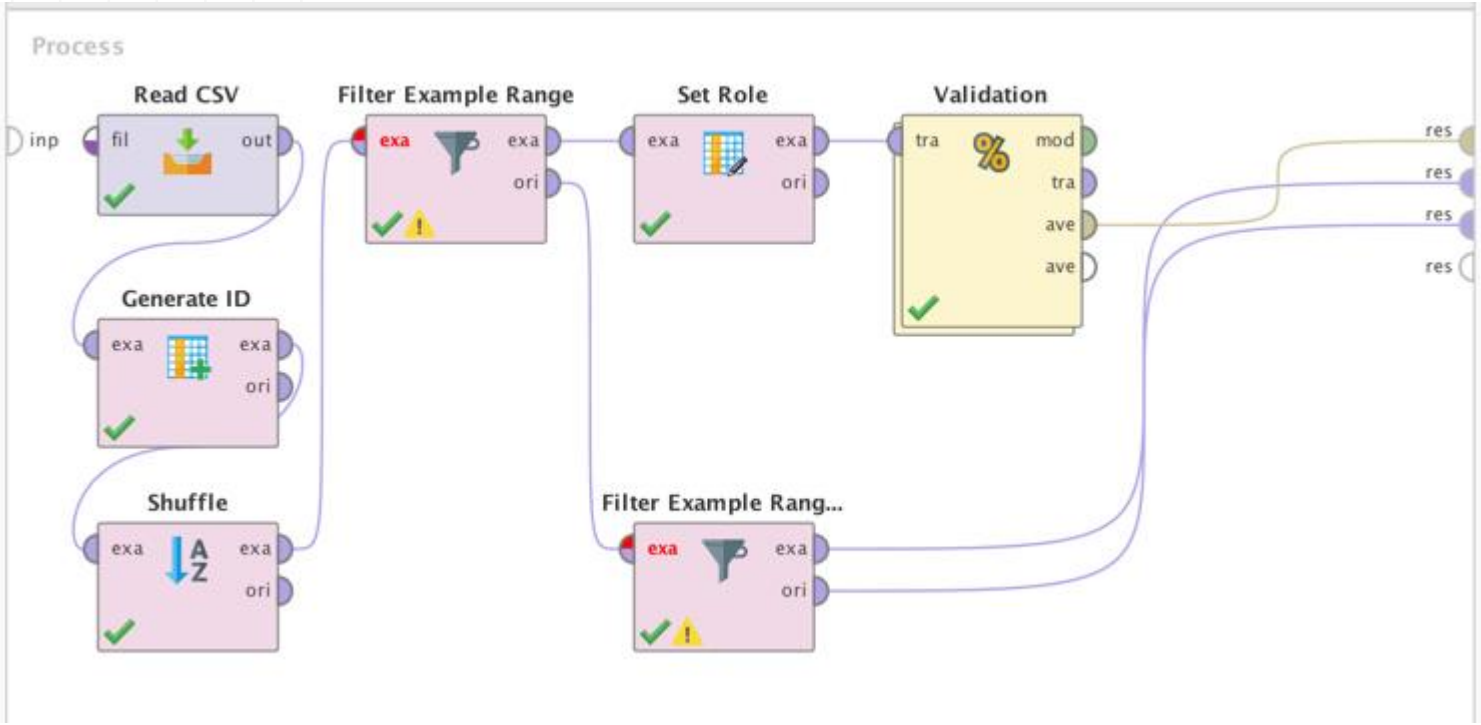
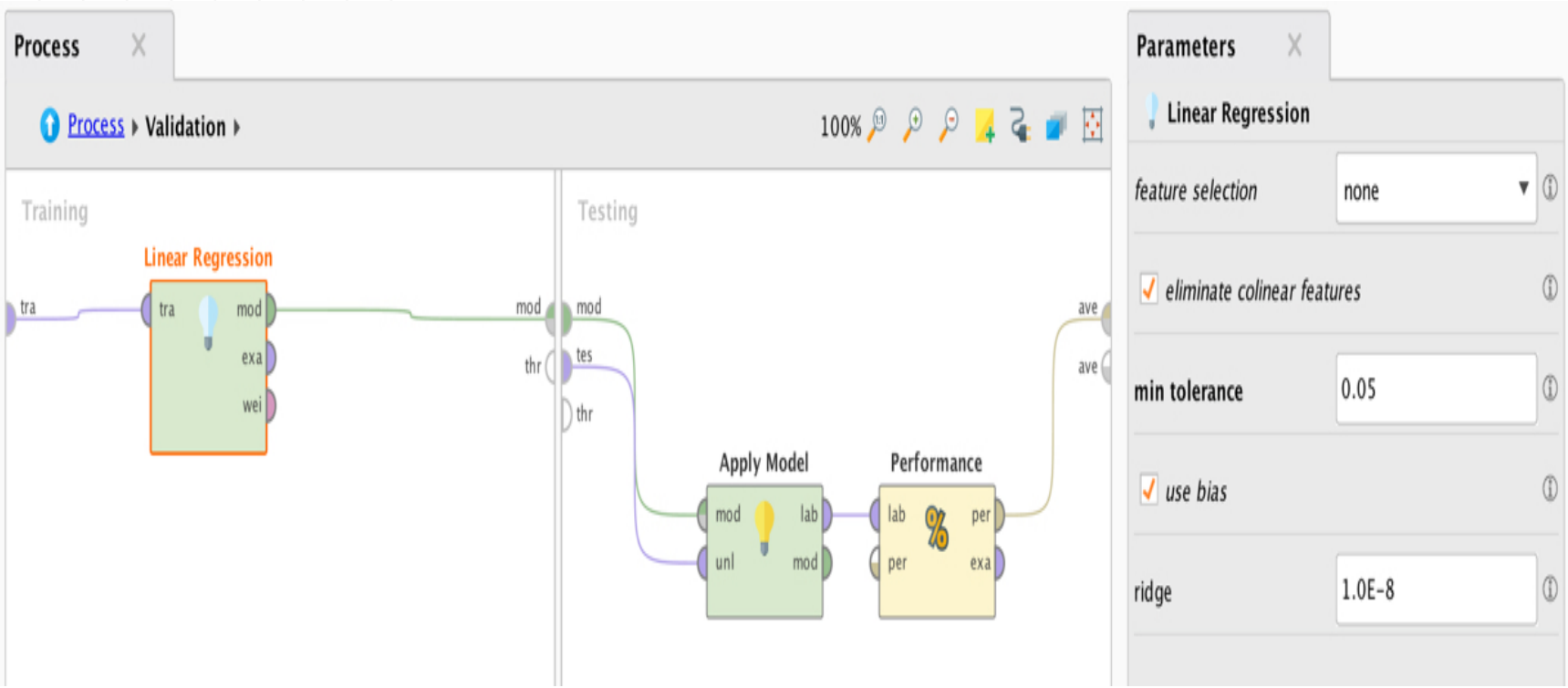
Chapter 5. Regression

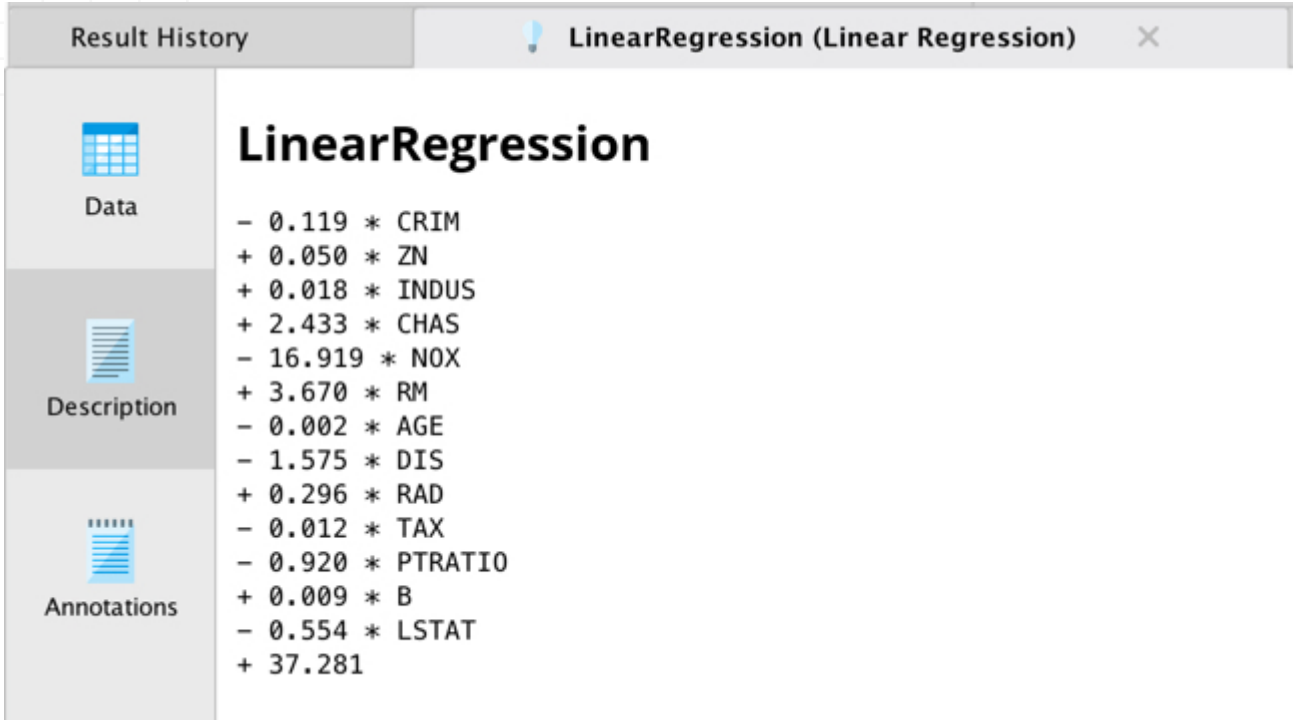


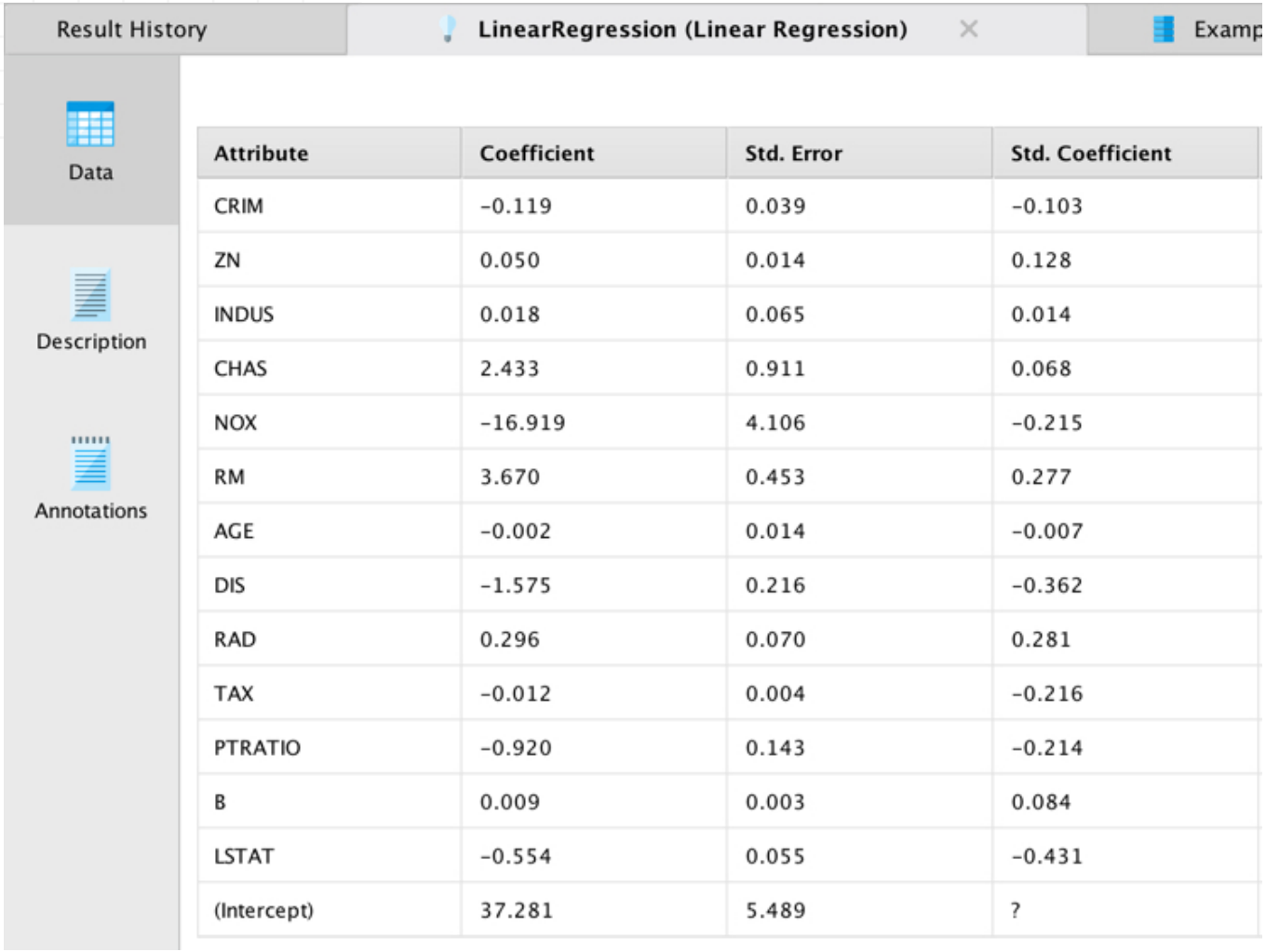

Median Home Price in '000s
60
50
40
30
20
10
0
0
1
2
3
4
5
6
7
8
9
10
Number of Rooms


Process
Read CSV
Filter Example Range
inp
fil
out
exa
exa
ori
res
res
res
Generate ID
exa
exa
ori
Shuffle
exa
exa
ori
Filter Example Rang...
exa
exa
ori


Process
Read CSV
inp
fil
out
Filter Example Range
exa
exa
ori
Set Role
exa
exa
ori
Validation
tra
mod
tra
ave
ave
res
res
res
res
Generate ID
exa
exa
ori
Shuffle
exa
exa
ori
Filter Example Rang...
exa
exa
ori


Process
Process ▸ Validation ▸
100%
Training
Linear Regression
tra
tra
mod
exa
wei
mod
thr
Testing
mod
tes
thr
Apply Model
mod
lab
unl
mod
Performance
lab
per
per
exa
ave
ave
Parameters
Linear Regression
feature selection
none
eliminate colinear features
min tolerance
0.05
use bias
ridge
1.0E-8

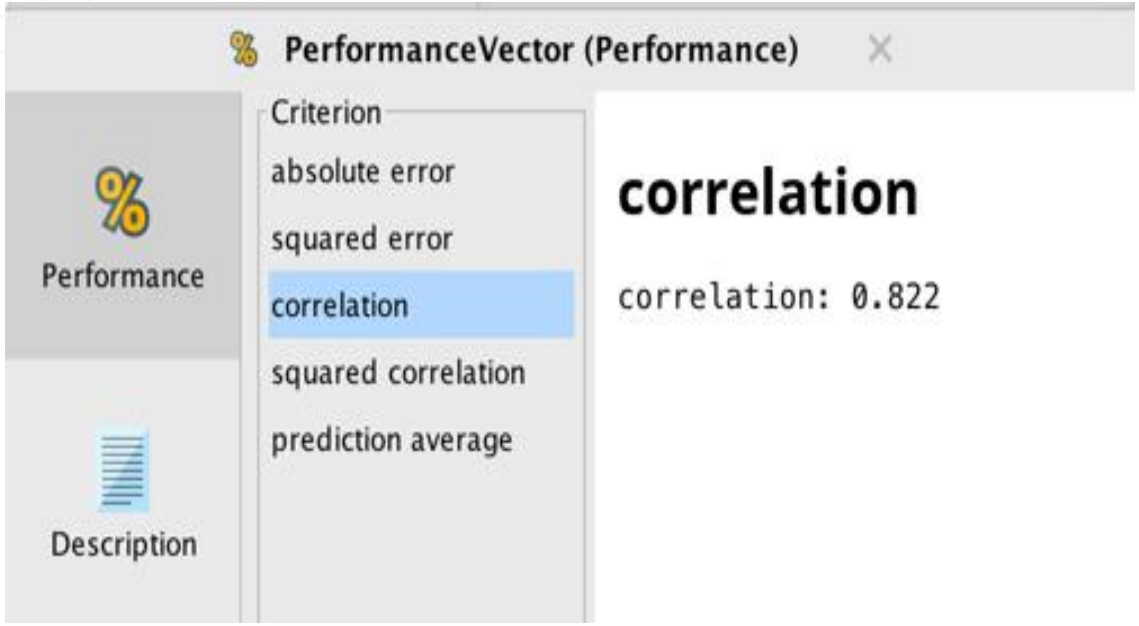

Result History
LinearRegression (Linear Regression)
Data
Description
Annotations
LinearRegression
- 0.119 * CRIM
+ 0.050 * ZN
+ 0.018 * INDUS
+ 2.433 * CHAS
- 16.919 * NOX
+ 3.670 * RM
- 0.002 * AGE
- 1.575 * DIS
+ 0.296 * RAD
- 0.012 * TAX
- 0.920 * PTRATIO
+ 0.009 * B
- 0.554 * LSTAT
+ 37.281



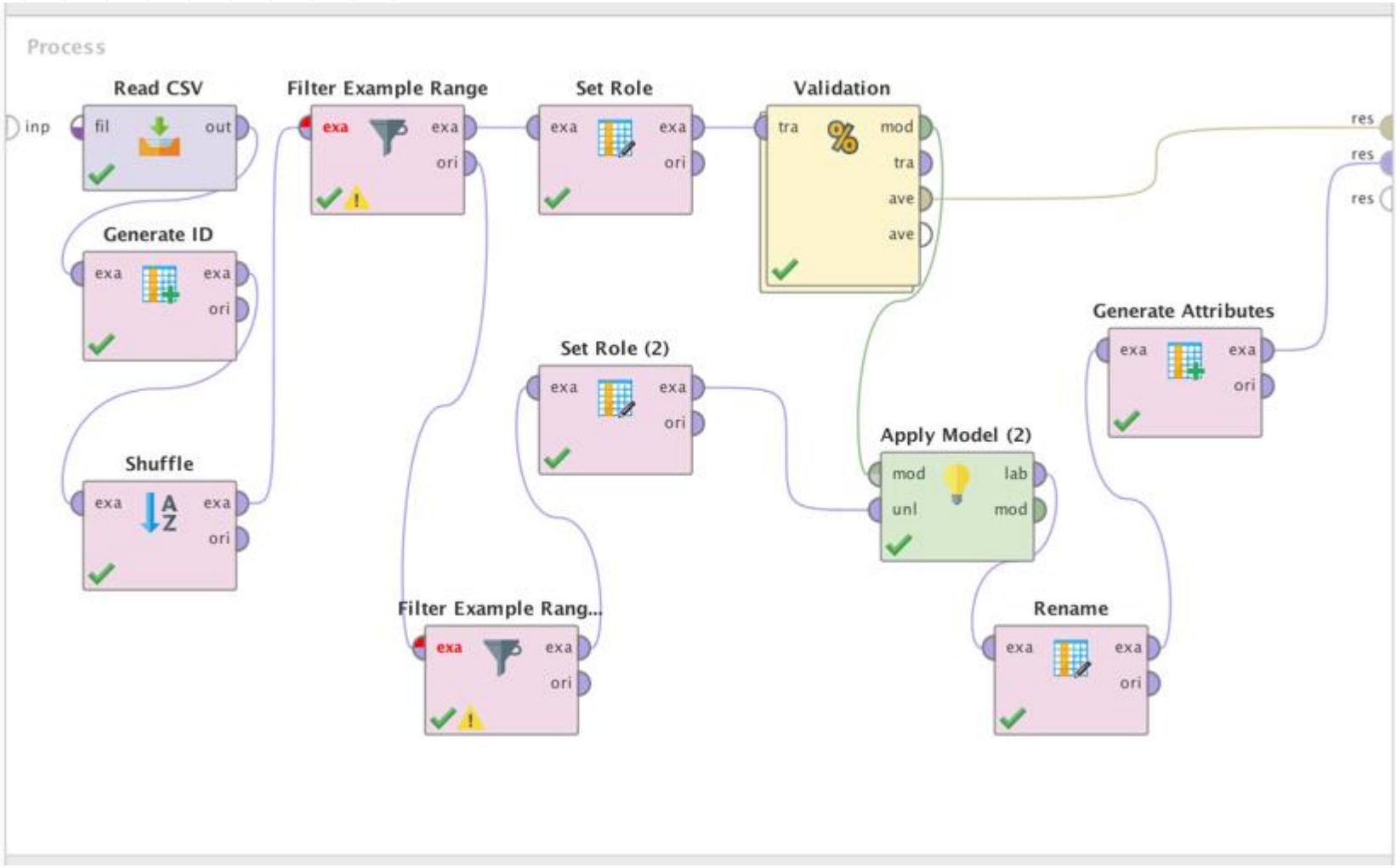
Attribute	Coefficient	Std. Error	Std. Coefficient
CRIM	-0.119	0.039	-0.103
ZN	0.050	0.014	0.128
INDUS	0.018	0.065	0.014
CHAS	2.433	0.911	0.068
NOX	-16.919	4.106	-0.215
RM	3.670	0.453	0.277
AGE	-0.002	0.014	-0.007
DIS	-1.575	0.216	-0.362
RAD	0.296	0.070	0.281
TAX	-0.012	0.004	-0.216
PTRATIO	-0.920	0.143	-0.214
B	0.009	0.003	0.084
LSTAT	-0.554	0.055	-0.431
(Intercept)	37.281	5.489	?

LinearRegression

```
- 0.119 * CRIM
+ 0.049 * ZN
+ 2.444 * CHAS
- 16.784 * NOX
+ 3.646 * RM
- 1.579 * DIS
+ 0.292 * RAD
- 0.011 * TAX
- 0.916 * PTRATIO
+ 0.009 * B
- 0.556 * LSTAT
+ 37.258
```


PerformanceVector (Performance)
Performance
Description
Criterion
absolute error
squared error
correlation
squared correlation
prediction average
correlation
correlation: 0.822

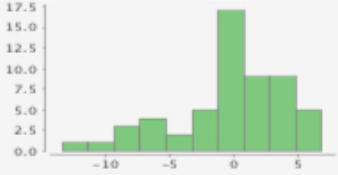
Attribute	Coefficient	Std. Error	Std. Coeff...	Tolerance	t-Stat	p-Value ↑	Code
LSTAT	-0.556	0.052	-0.432	0.490	-10.661	0	****
RM	3.646	0.443	0.275	0.581	8.236	0.000	****
DIS	-1.579	0.199	-0.363	0.823	-7.928	0.000	****
(Intercept)	37.258	5.440	?	?	6.849	0.000	****
PTRATIO	-0.916	0.140	-0.213	0.793	-6.547	0.000	****
NOX	-16.784	3.794	-0.213	0.812	-4.424	0.000	****
RAD	0.292	0.068	0.276	0.769	4.318	0.000	****
ZN	0.049	0.014	0.128	0.877	3.465	0.001	****
TAX	-0.011	0.004	-0.208	0.749	-3.219	0.001	***
CRIM	-0.119	0.039	-0.103	0.843	-3.088	0.002	***
B	0.009	0.003	0.083	0.905	2.953	0.003	***
CHAS	2.444	0.905	0.068	0.991	2.701	0.007	***

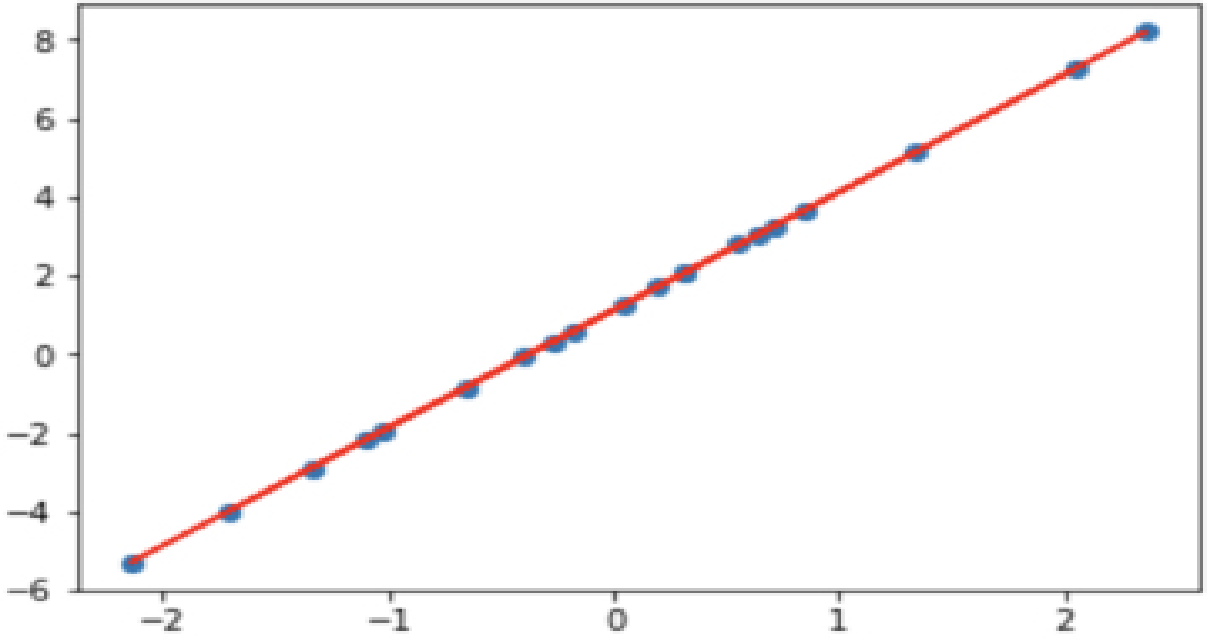

Process
Read CSV
inp
fil
out
Filter Example Range
exa
exa
ori
Set Role
exa
exa
ori
Validation
tra
mod
tra
ave
ave
res
res
res
Generate ID
exa
exa
ori
Shuffle
exa
exa
ori
Set Role (2)
exa
exa
ori
Filter Example Rang...
exa
exa
ori
Apply Model (2)
mod
lab
unl
mod
Generate Attributes
exa
exa
ori
Rename
exa
exa
ori

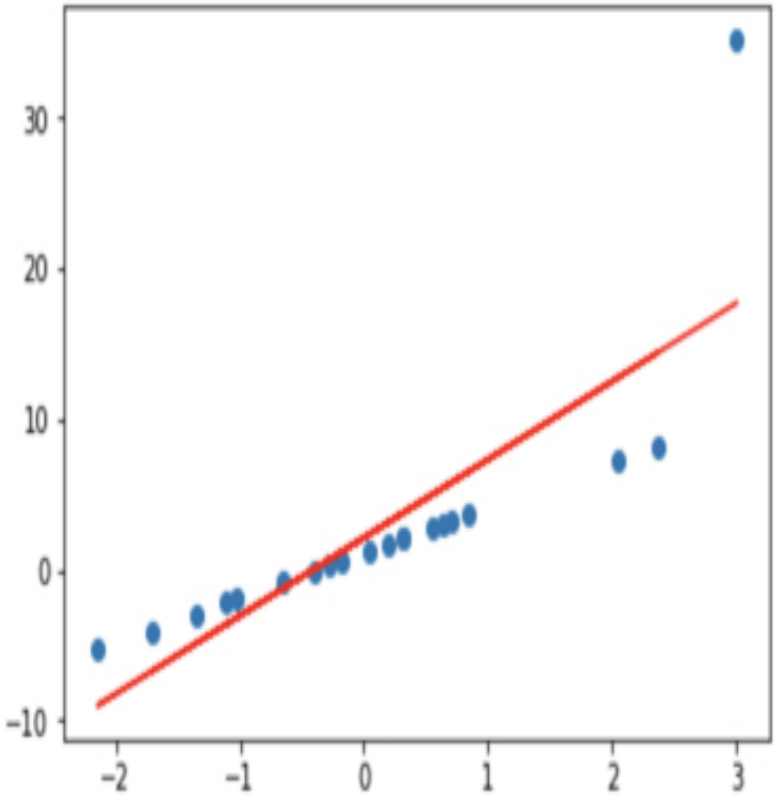
Result History | ExampleSet (Generate Attributes) | PerformanceVector (Performance)

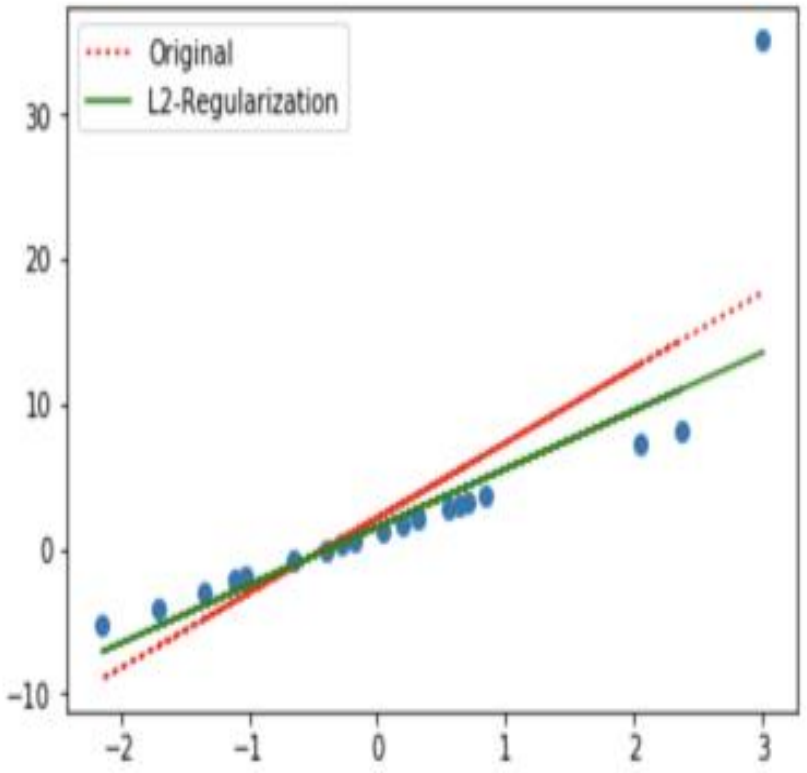
Data
Statistics
Charts
Advanced Charts
Annotations

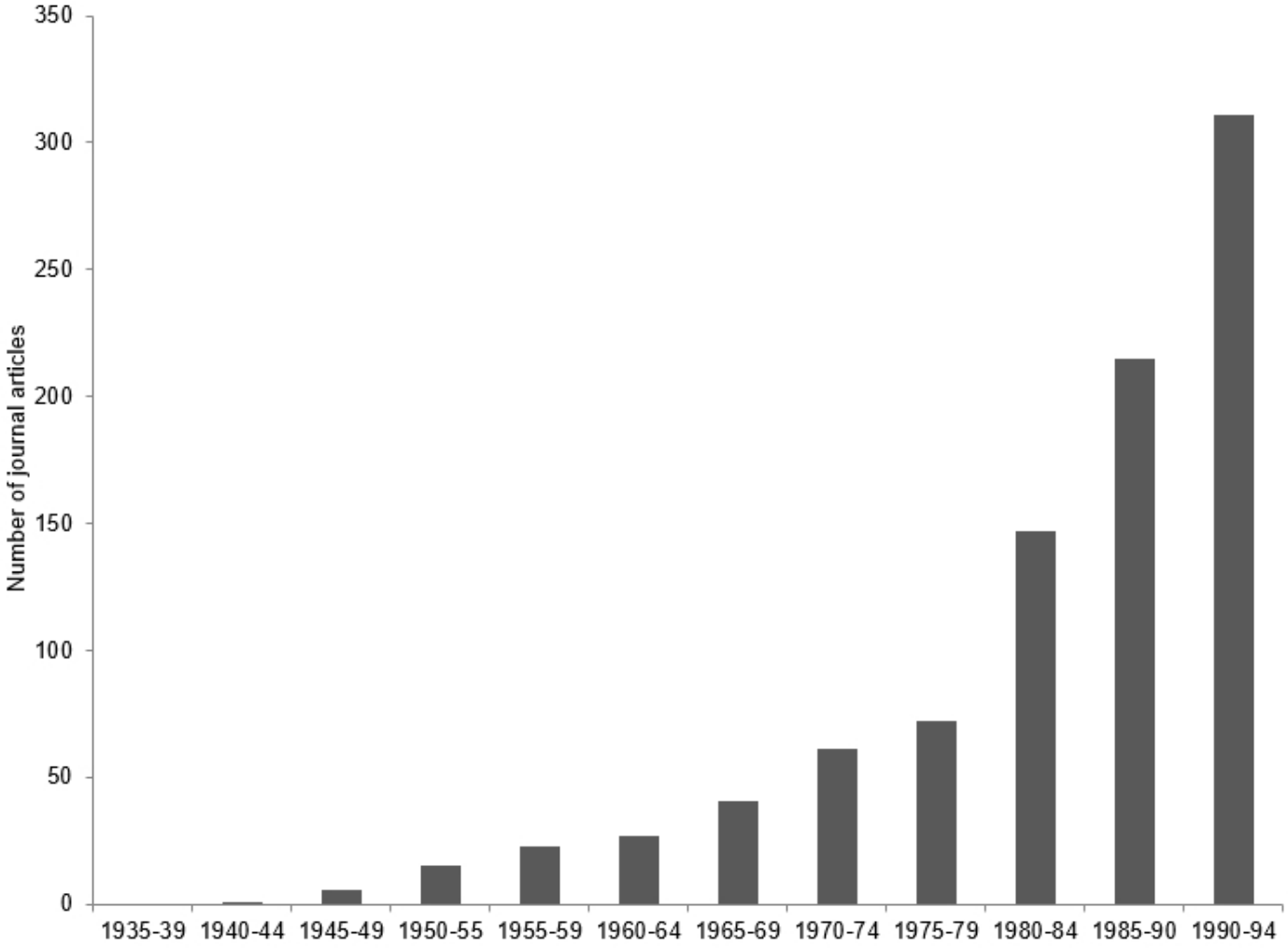
Filter (17 / 17 attributes): Search for Attribute.

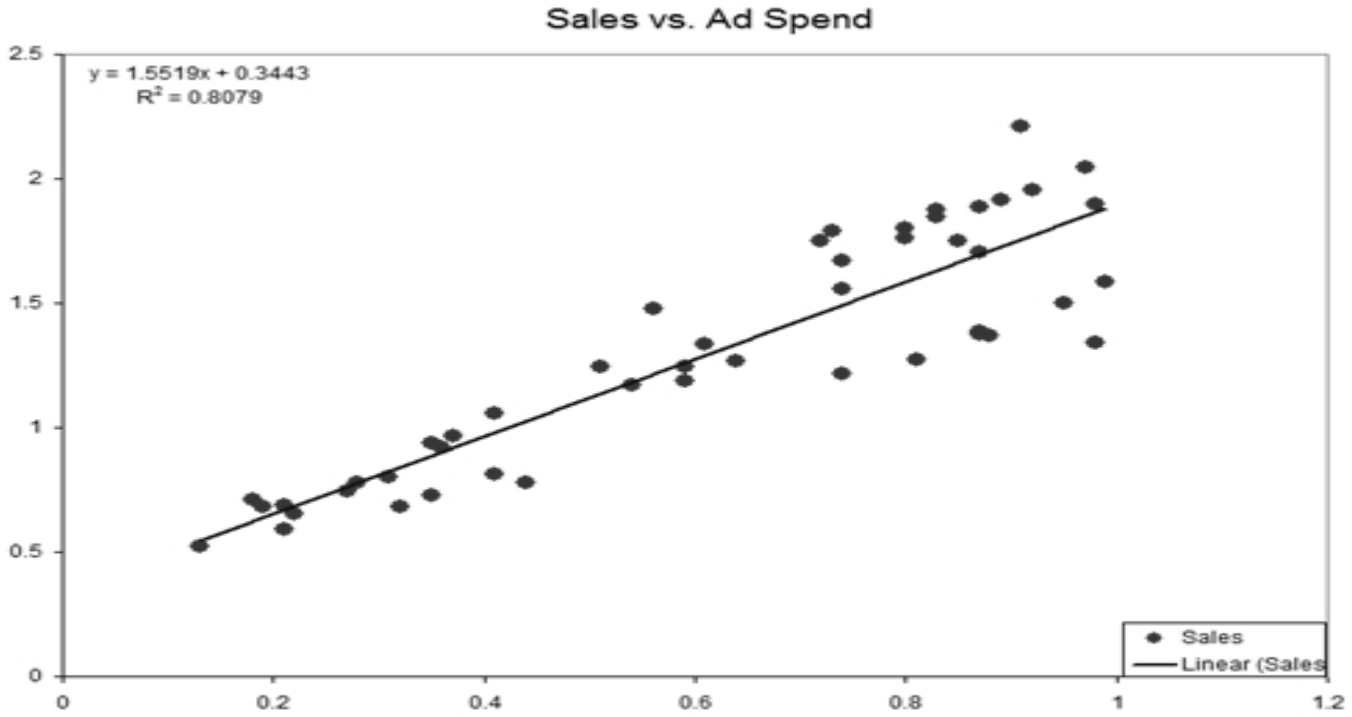
Name	Type	Missing	Statistics		
RAD	Real	0	Min 1	Max 24	Average 9.893
TAX	Real	0	Min 193	Max 666	Average 414.179
PTRATIO	Real	0	Min 13	Max 21.200	Average 18.338
B	Real	0	Min 0.320	Max 396.900	Average 340.030
LSTAT	Real	0	Min 3.130	Max 37.970	Average 12.848
residuals	Real	0	 Open chart	Min -13.353	Max 6.763


8
6
4
2
0
−2
−4
−6
−2
−1
0
1
2

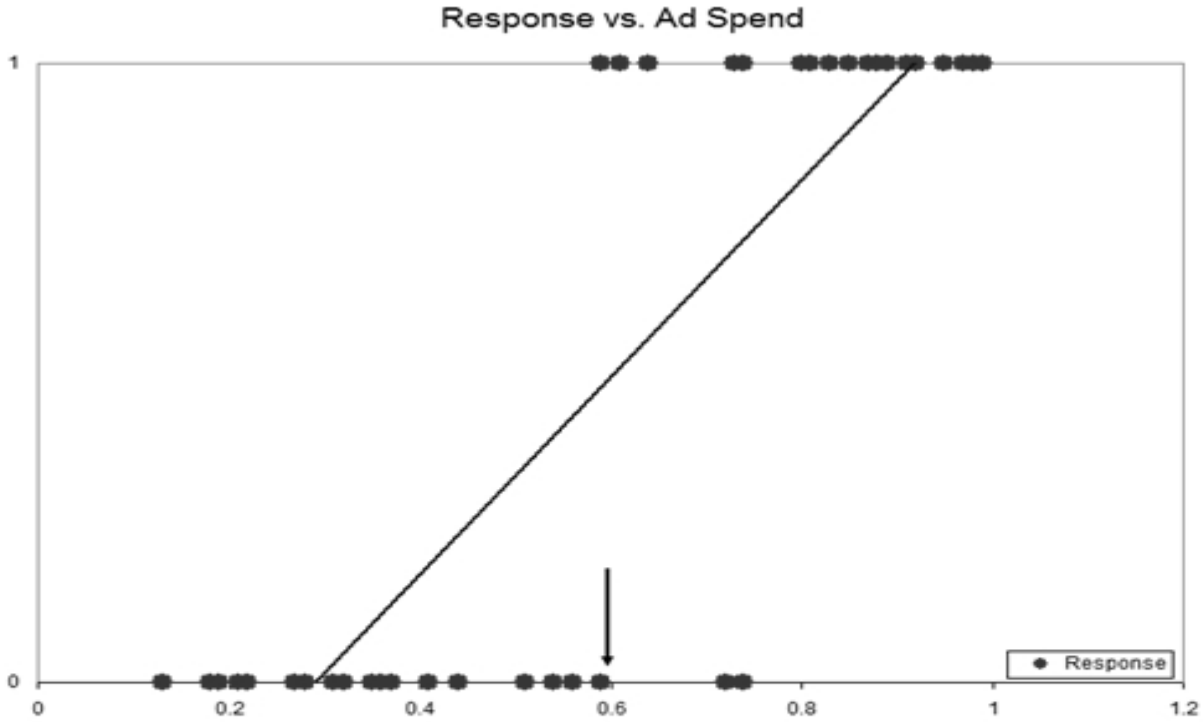

30
20
10
0
−10
−2
−1
0
1
2
3


Original
L2-Regularization
30
20
10
0
−10
−2
−1
0
1
2
3

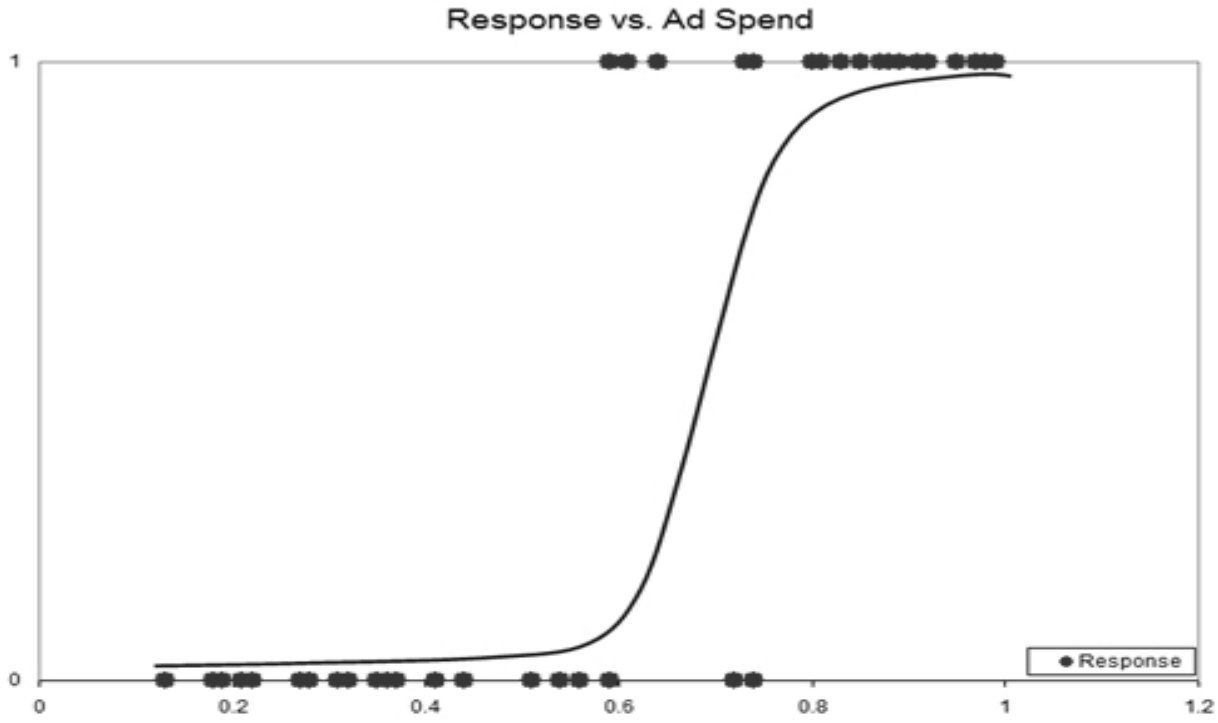

Number of journal articles
350
300
250
200
150
100
50
0
1935-39
1940-44
1945-49
1950-55
1955-59
1960-64
1965-69
1970-74
1975-79
1980-84
1985-90
1990-94



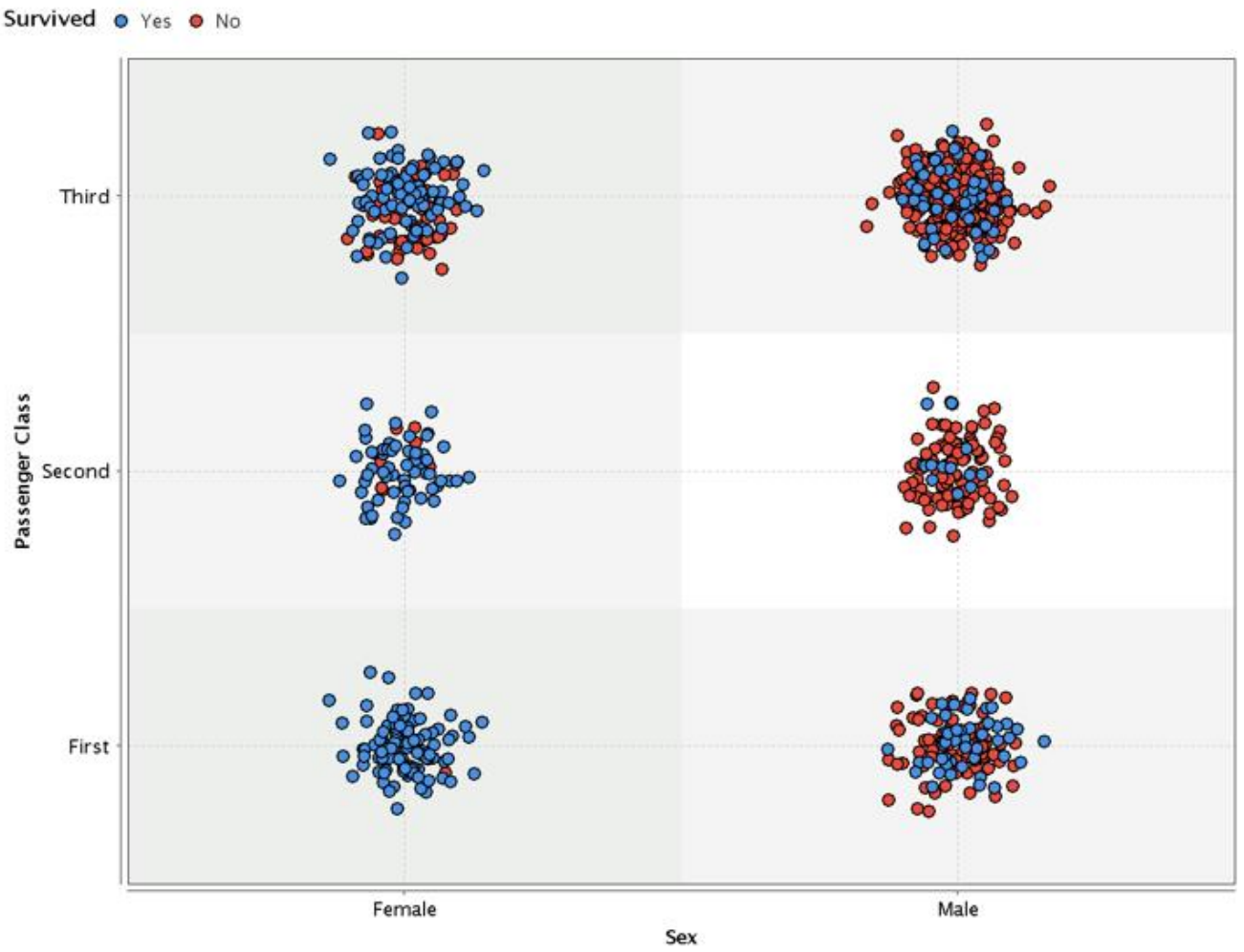
Linear Regression Model. We can make an intuitive assessment that increase in *Ad spend* also increases *Sales*. Using the straight line, we may also be able to predict.

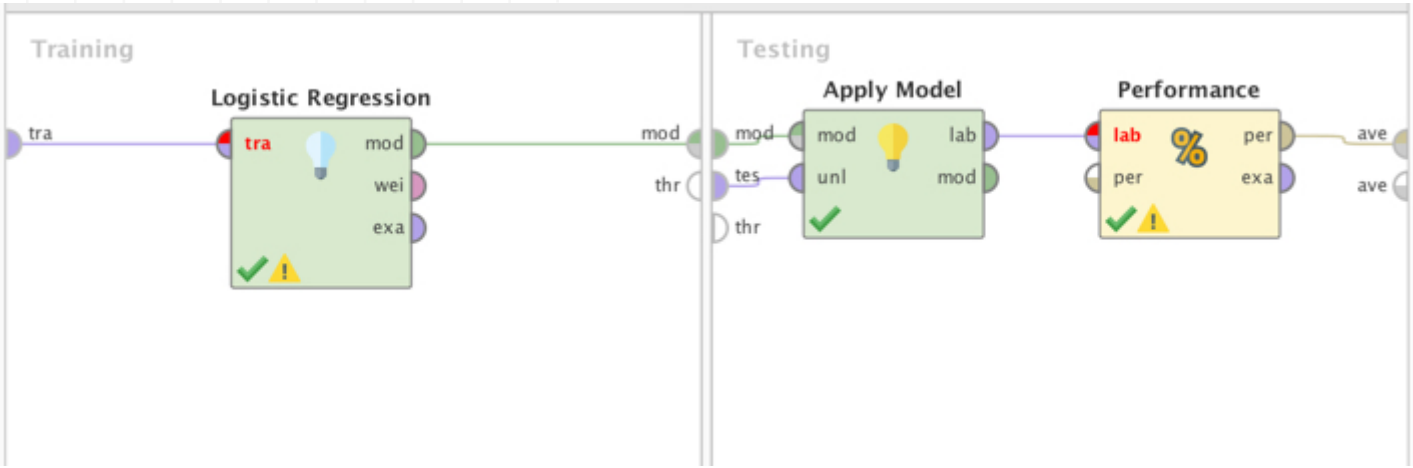


Linear Fit for a Binary outcome: Although we can make an intuitive assessment that increase in *Ad spend* increases *Response,* the switch is abrupt – around 0.6. Using the straight line, we cannot really predict outcome.



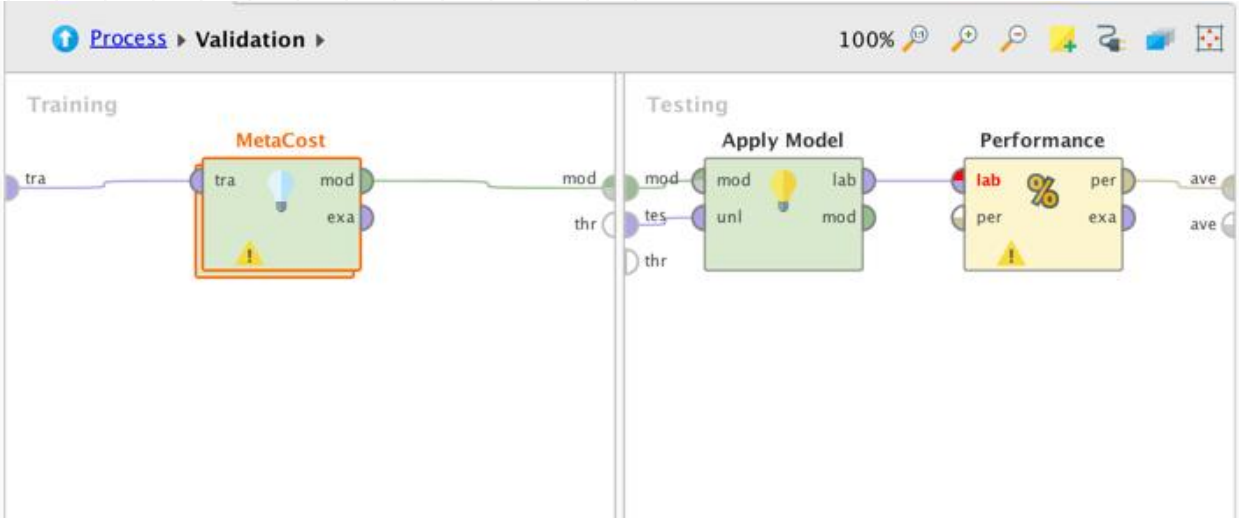
Logistic Regression Model. The S-shaped curve is clearly a better fit for *most* of the data. We can state *Ad spend* increases *Sales,* **and** we may also be able to predict using this model.

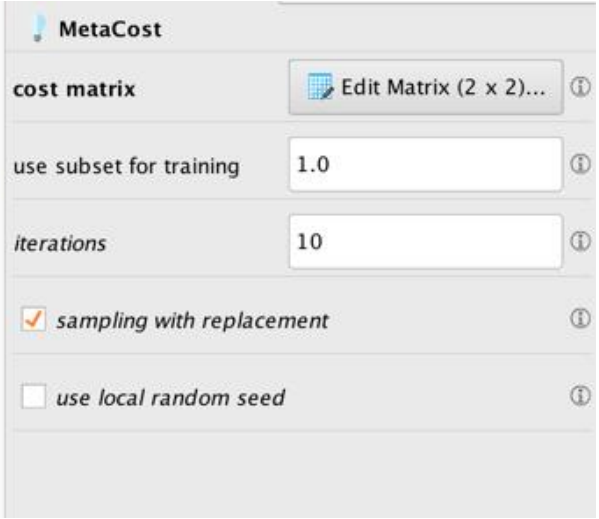

Survived
Yes
No
Third
Second
First
Passenger Class
Female
Male
Sex



accuracy: 83.33%

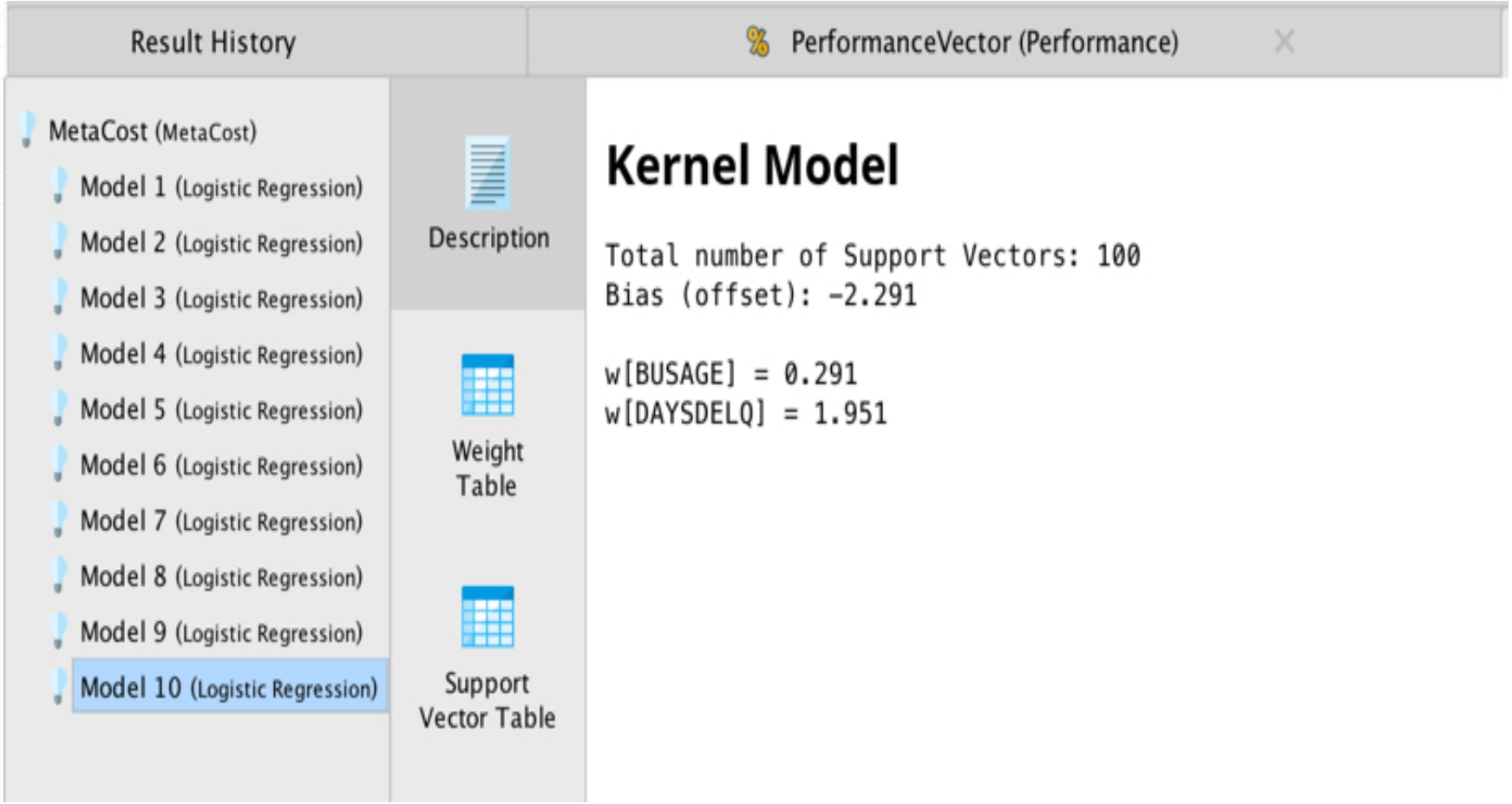
	true N	true Y	class precision
pred. N	21	3	87.50%
pred. Y	2	4	66.67%
class recall	91.30%	57.14%	





accuracy: 83.33%

	true N	true Y	class precision
pred. N	20	2	90.91%
pred. Y	3	5	62.50%
class recall	86.96%	71.43%	


Result History
PerformanceVector (Performance)
MetaCost (MetaCost)
Model 1 (Logistic Regression)
Model 2 (Logistic Regression)
Model 3 (Logistic Regression)
Model 4 (Logistic Regression)
Model 5 (Logistic Regression)
Model 6 (Logistic Regression)
Model 7 (Logistic Regression)
Model 8 (Logistic Regression)
Model 9 (Logistic Regression)
Model 10 (Logistic Regression)
Description
Weight Table
Support Vector Table
Kernel Model
Total number of Support Vectors: 100
Bias (offset): -2.291
w[BUSAGE] = 0.291
w[DAYSDELQ] = 1.951