Lesson 22: Feature Scoring and Selection

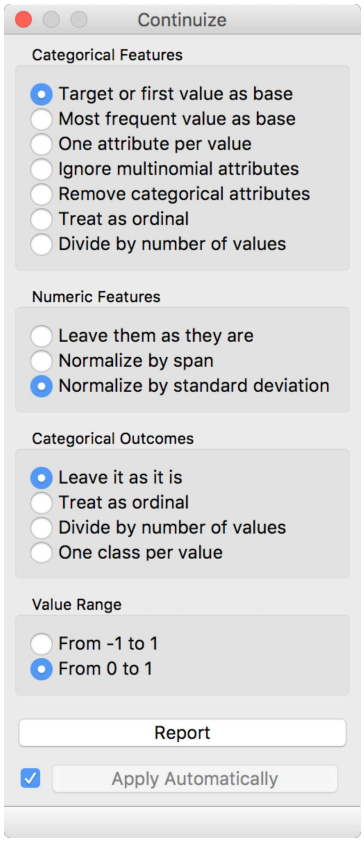
For this lesson, load the data from imports-85.tab using the File widget and Browse documentation data sets.

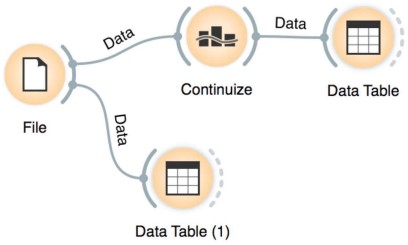
Linear regression infers a model that estimate the class, a real-valued feature, as a sum of products of input features and their weights. Consider the data on prices of imported cars in 1985.

Data Table (1)

	height	curb-weight	engine-type	num-of-cylinders	engine-size	fuel-system	bore	stroke
1	48.800	2548.000	dohc	four	130.000	mpfi	3.470	2.680
2	48.800	2548.000	dohc	four	130.000	mpfi	3.470	2.680
3	52.400	2823.000	ohcv	six	152.000	mpfi	2.680	3.470
4	54.300	2337.000	ohc	four	109.000	mpfi	3.190	3.400
5	54.300	2824.000	ohc	five	136.000	mpfi	3.190	3.400
6	53.100	2507.000	ohc	five	136.000	mpfi	3.190	3.400
7	55.700	2844.000	ohc	five	136.000	mpfi	3.190	3.400
8	55.700	2954.000	ohc	five	136.000	mpfi	3.190	3.400
9	55.900	3086.000	ohc	five	131.000	mpfi	3.130	3.400
10	52.000	3053.000	ohc	five	131.000	mpfi	3.130	3.400
11	54.300	2395.000	ohc	four	108.000	mpfi	3.500	2.800
12	54.300	2395.000	ohc	four	108.000	mpfi	3.500	2.800
13	54.300	2710.000	ohc	six	164.000	mpfi	3.310	3.190
14	54.300	2765.000	ohc	six	164.000	mpfi	3.310	3.190
15	55.700	3055.000	ohc	six	164.000	mpfi	3.310	3.190
16	55.700	3230.000	ohc	six	209.000	mpfi	3.620	3.390
17	53.700	3380.000	ohc	six	209.000	mpfi	3.620	3.390
18	56.300	3505.000	ohc	six	209.000	mpfi	3.620	3.390

Inspecting this data set in a Data Table, it shows that some features, like fuel-system, engine-type and many others, are discrete. Linear regression only works with numbers. In Orange, linear regression will automatically convert all discrete values to numbers, most often using several features to represent a single discrete features. We also do this conversion manually by using Continuize widget.



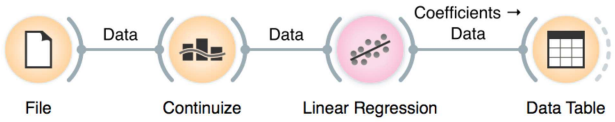


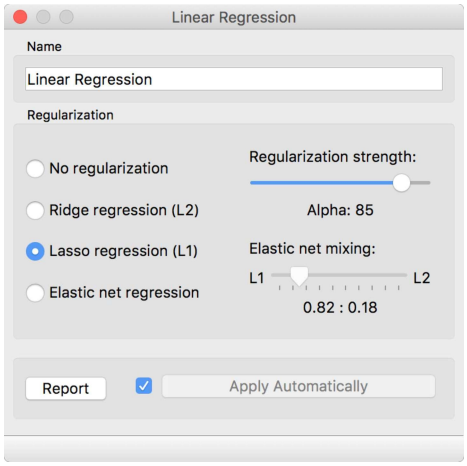
Before we continue, you should check what Continuize actually does and how it converts the nominal features into real-valued features. The table below should provide sufficient illustration.

Data Table

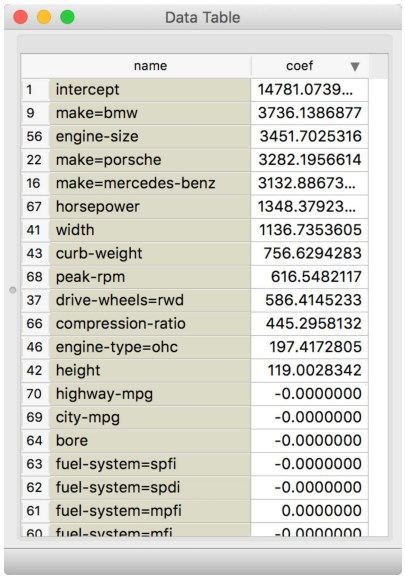
	symboling=3	ormalized-losses	make=audi	make=bmw	make=chevrolet	make=dodge
1	1.000	?	0.000	0.000	0.000	0.000
2	1.000	?	0.000	0.000	0.000	0.000
3	0.000	?	0.000	0.000	0.000	0.000
4	0.000	1.189	1.000	0.000	0.000	0.000
5	0.000	1.189	1.000	0.000	0.000	0.000
6	0.000	?	1.000	0.000	0.000	0.000
7	0.000	1.019	1.000	0.000	0.000	0.000
8	0.000	?	1.000	0.000	0.000	0.000
9	0.000	1.019	1.000	0.000	0.000	0.000
10	0.000	?	1.000	0.000	0.000	0.000
11	0.000	1.981	0.000	1.000	0.000	0.000
12	0.000	1.981	0.000	1.000	0.000	0.000
13	0.000	1.868	0.000	1.000	0.000	0.000
14	0.000	1.868	0.000	1.000	0.000	0.000
15	0.000	?	0.000	1.000	0.000	0.000
16	0.000	?	0.000	1.000	0.000	0.000
17	0.000	?	0.000	1.000	0.000	0.000
18	0.000	?	0.000	1.000	0.000	0.000
19	0.000	-0.028	0.000	0.000	1.000	0.000
20	0.000	-0.679	0.000	0.000	1.000	0.000

Now to the core of this lesson. Our workflow reads the data, coninuizes it such that we also normalize all the features to bring them the to equal scale, then we load the data into Linear Regression widget and check out the feature coefficients in the Data Table.





In Linear Regression, we will use L1 regularization. Compared to L2 regularization, which aims to minimize the sum of squared weights, L1 regularization is more rough and minimizes the sum of absolute values of the weights. The result of this "roughness" is that many of the feature will get zero weights.


Data Table

	name	coef
1	intercept	14781.0739...
9	make=bmw	3736.1386877
56	engine-size	3451.7025316
22	make=porsche	3282.1956614
16	make=mercedes-benz	3132.88673...
67	horsepower	1348.37923...
41	width	1136.7353605
43	curb-weight	756.6294283
68	peak-rpm	616.5482117
37	drive-wheels=rwd	586.4145233
66	compression-ratio	445.2958132
46	engine-type=ohc	197.4172805
42	height	119.0028342
70	highway-mpg	-0.0000000
69	city-mpg	-0.0000000
64	bore	-0.0000000
63	fuel-system=spfi	-0.0000000
62	fuel-system=spdi	-0.0000000
61	fuel-system=mpfi	0.0000000
60	fuel-system=mfi	-0.0000000

But this may be also exactly what we want. We want to select only the most important features, and want to see how the model that uses only a smaller subset of features actually behaves. Also, this smaller set of features is ranked. Engine size is a huge factor in pricing of our cars, and so is the make, where Porsche, Mercedes and BMW cost more than other cars (ok, no news here).

We should notice that the number of features with non-zero weights varies with regularization strength. Stronger regularization would result in fewer features with non-zero weights.