Joint Data Analysis

Working notes for the hands-on course at Procter & Gamble

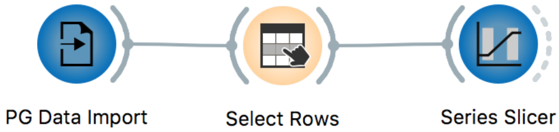
This tutorial will combine data from the previous days — measured instrument data and the consumer study data — and try to build predictive models on it.

Handouts prepared by:
Bruna Zupan, Lan Žagar and Primož Godec

Lesson 1: DRF

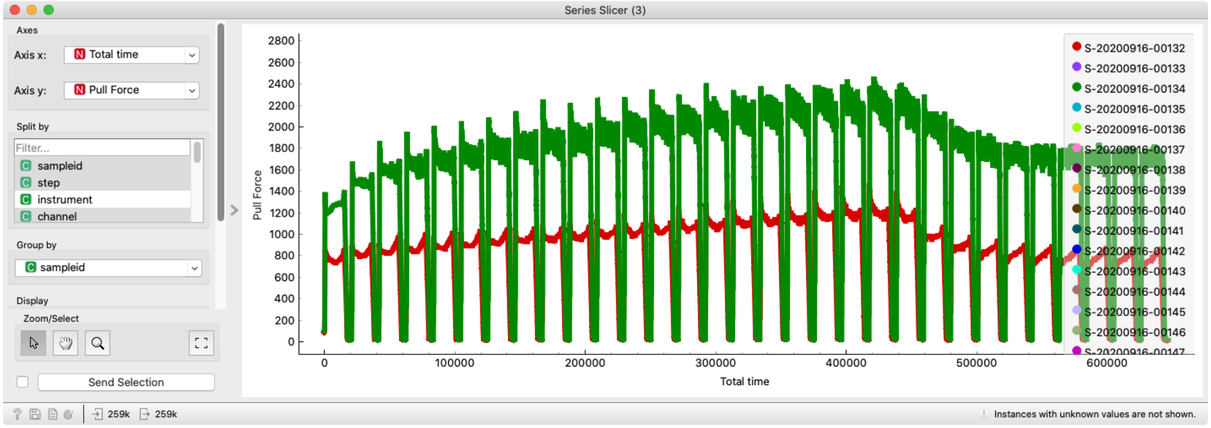
This lesson will explore DRF data and its possible relationship with selected questions from the Consumer study data set.

In *PG Data Import*, we load all DRF data, Metadata, and Consumer study data. Note that the DRF experiment has a much bigger data set than the rest of the instruments (6.6 million measurements).



To first visualize a small sample, connect DRF data with the *Select Rows* widget and select two sample ids and step 8. Then, use the *Series Slicer* widget to visualize them.

For Axis x, select `Total Time` to plot all pulls sequentially or `Time` to plot all pulls simultaneously (overlapping). For Axis y, select `Pull Force`. In Split by select `sampleid`, `step`, and `channel`.


Series Slicer (3)

Axes
Axis x: Total time
Axis y: Pull Force

Split by
sampleid
step
instrument
channel

Group by
sampleid

Display
Zoom/Select

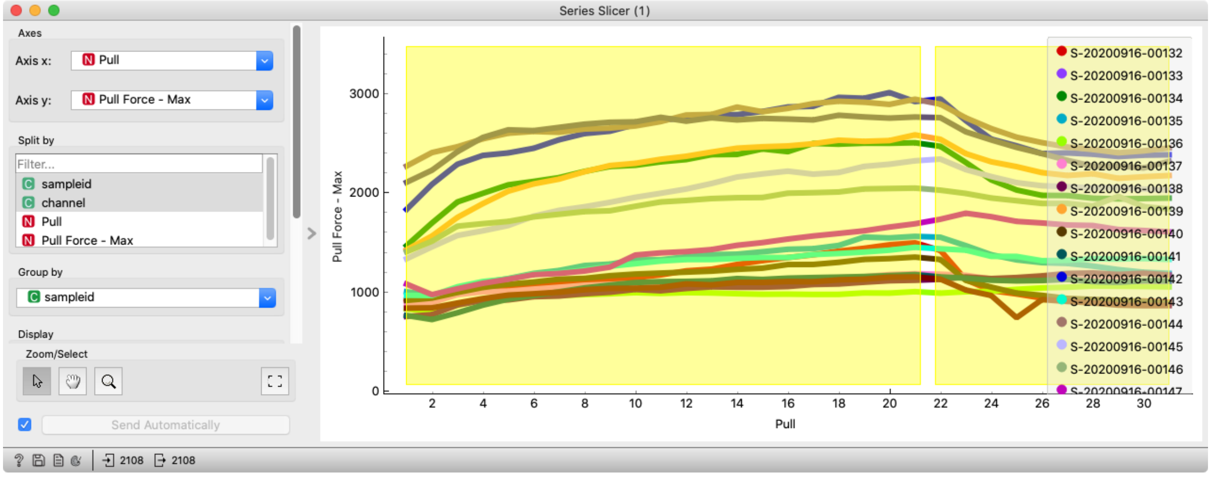
□ Send Selection

259k 259k

Instances with unknown values are not shown.

Going back to all data, use *Select Rows* to select just step 8 and all samples. Then use the *Aggregate* widget to aggregate each pull with the max `Pull Force` to obtain a summarized series of 31 pull forces.

Connect *Aggregate* widget to the *Series Slicer* widget to visualize aggregated data as a time series. Then construct two slices: shampoo phase (pull 1-21) and rinse phase (pull 22-31).


Series Slicer (1)

Axes

Axis x: Pull

Axis y: Pull Force - Max

Split by

sampleid
channel
Pull
Pull Force - Max

Group by: sampleid

Display

Zoom/Select

☑ Send Automatically

2108 2108

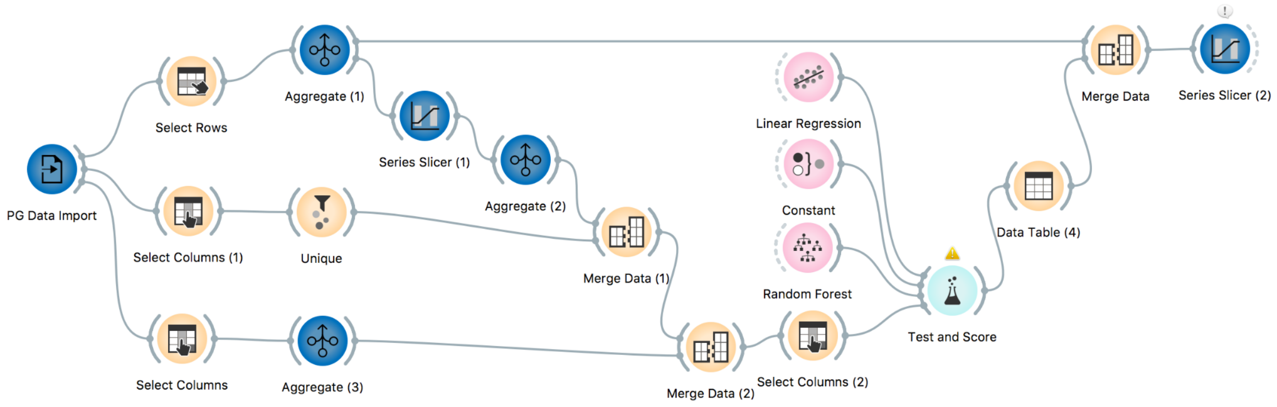
Use another *Aggregate* widget to model the `Pull Force - Max` trends in each slice with a quadratic function and summarize the overall forces with the area under the curve.

Use Metadata from *PG Data Import* to add batch information to the DRF data. First, use *Select Columns* to select `sampleid`, `batchid`, `batchdesc`, and `productdesc` from Metadata then connect it to the *Unique* widget to delete repeated instances. Then use the *Merge* widget to append extra columns to the main DRF (on Data channel) matching rows by `sampleid`.

Connect the *PG Data Import* widget to the *Select Columns* widget with the Consumer Study data. Select the question “Hair feels smooth while rinsing shampoo”. Then use *Aggregate* to compute the average rating for each batch. Merge the ratings to DRF data.

Now we have prepared the dataset to explore the relationship between DFR measurements and the consumer's answers.

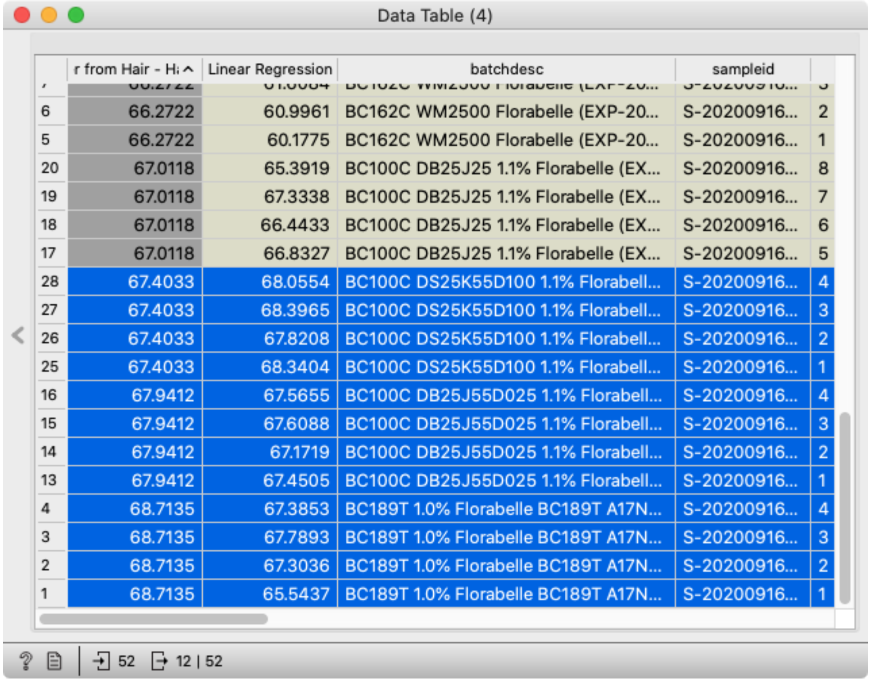
Note that Orange workflow for exploring the DRF and Consumer study data is available among the training documentation. Below are highlights and hints on how to start the investigation and what to observe.



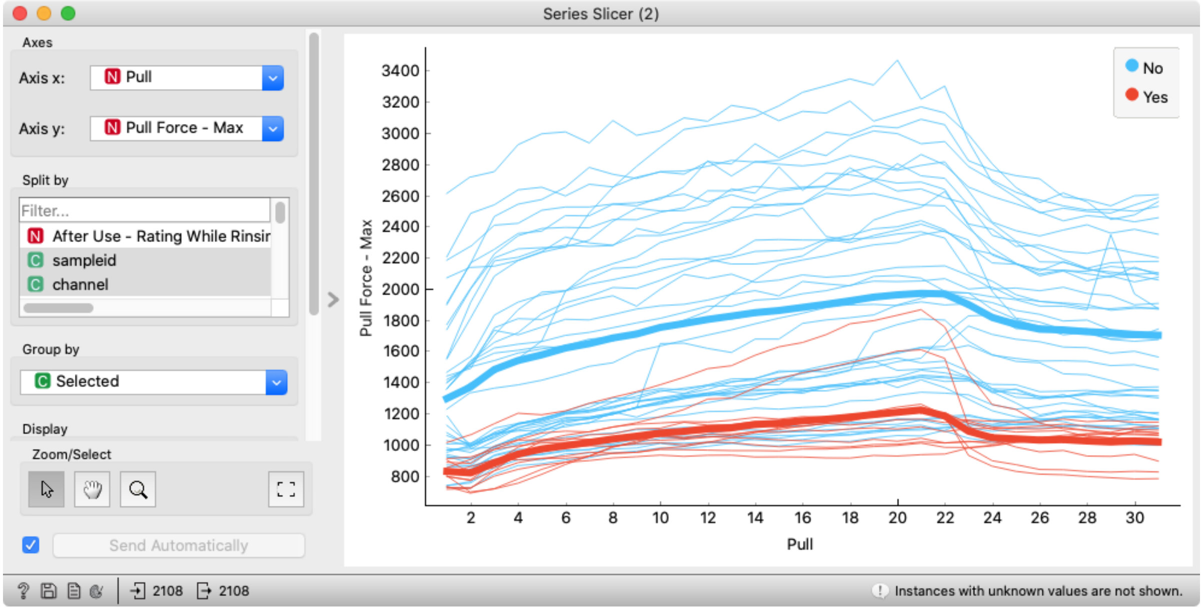
Train some models that can predict the consumer ratings from the profiled DRF data and assess their accuracy.

Instead of the standard cross-validation, try using Cross-validation by feature (in this example use `batchid`). We should use this type of cross-validation to exclude overfitting due to having the same sample tested multiple times.

Check the predicted vs. true ratings and plot the profiles for some selected examples.



	r from Hair - Hi	Linear Regression	batchdesc	sampleid	
7	66.2722	61.0084	BC162C WM2500 Florabelle (EXP-20...	S-20200916...	3
6	66.2722	60.9961	BC162C WM2500 Florabelle (EXP-20...	S-20200916...	2
5	66.2722	60.1775	BC162C WM2500 Florabelle (EXP-20...	S-20200916...	1
20	67.0118	65.3919	BC100C DB25J25 1.1% Florabelle (EX...	S-20200916...	8
19	67.0118	67.3338	BC100C DB25J25 1.1% Florabelle (EX...	S-20200916...	7
18	67.0118	66.4433	BC100C DB25J25 1.1% Florabelle (EX...	S-20200916...	6
17	67.0118	66.8327	BC100C DB25J25 1.1% Florabelle (EX...	S-20200916...	5
28	67.4033	68.0554	BC100C DS25K55D100 1.1% Florabell...	S-20200916...	4
27	67.4033	68.3965	BC100C DS25K55D100 1.1% Florabell...	S-20200916...	3
26	67.4033	67.8208	BC100C DS25K55D100 1.1% Florabell...	S-20200916...	2
25	67.4033	68.3404	BC100C DS25K55D100 1.1% Florabell...	S-20200916...	1
16	67.9412	67.5655	BC100C DB25J55D025 1.1% Florabell...	S-20200916...	4
15	67.9412	67.6088	BC100C DB25J55D025 1.1% Florabell...	S-20200916...	3
14	67.9412	67.1719	BC100C DB25J55D025 1.1% Florabell...	S-20200916...	2
13	67.9412	67.4505	BC100C DB25J55D025 1.1% Florabell...	S-20200916...	1
4	68.7135	67.3853	BC189T 1.0% Florabelle BC189T A17N...	S-20200916...	4
3	68.7135	67.7893	BC189T 1.0% Florabelle BC189T A17N...	S-20200916...	3
2	68.7135	67.3036	BC189T 1.0% Florabelle BC189T A17N...	S-20200916...	2
1	68.7135	65.5437	BC189T 1.0% Florabelle BC189T A17N...	S-20200916...	1



Series Slicer (2)

Axes

Axis x: Pull

Axis y: Pull Force - Max

Split by

Filter...

After Use - Rating While Rinsir

sampleid

channel

Group by

Selected

Display

Zoom/Select

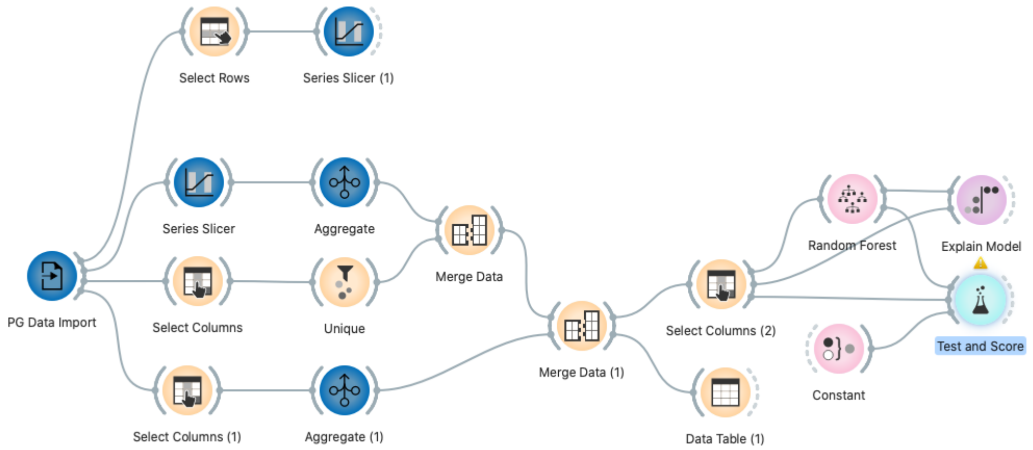
☑ Send Automatically

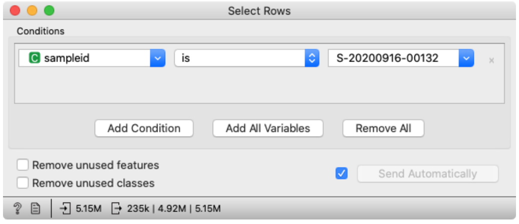
2108 2108

Instances with unknown values are not shown.

Lesson 2: Instron

In this lesson, we will load the Instron data, inspect them with the *Series Slicer*, merge them with the consumer study data, and predict the attributes from the consumer study. With the prediction model, we will observe which attributes from the Instron data contribute the most to predicting two combing-related consumer ratings.




Select Rows

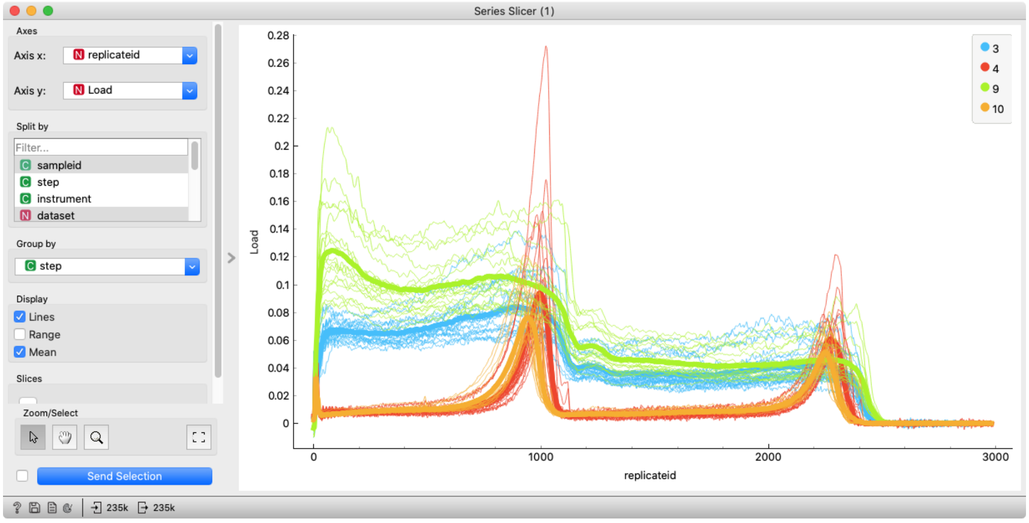
Conditions

sampleid is S-20200916-00132

□ Remove unused features
□ Remove unused classes
☑ Send Automatically

5.15M 235k | 4.92M | 5.15M

In *PG Data Import* we load all Instron data, Metadata, and Consumer study data. We then select one sample with *Select Rows* and inspect it in the *Series slicer (1)* widget. In the slicer, we can see that there are significant differences between steps 3, 9, and steps 4, 10.


Series Slicer (1)

Axes

Axis x: replicateid

Axis y: Load

Split by

Filter...
sampleid
step
instrument
dataset

Group by: step

Display

☑ Lines
□ Range
☑ Mean

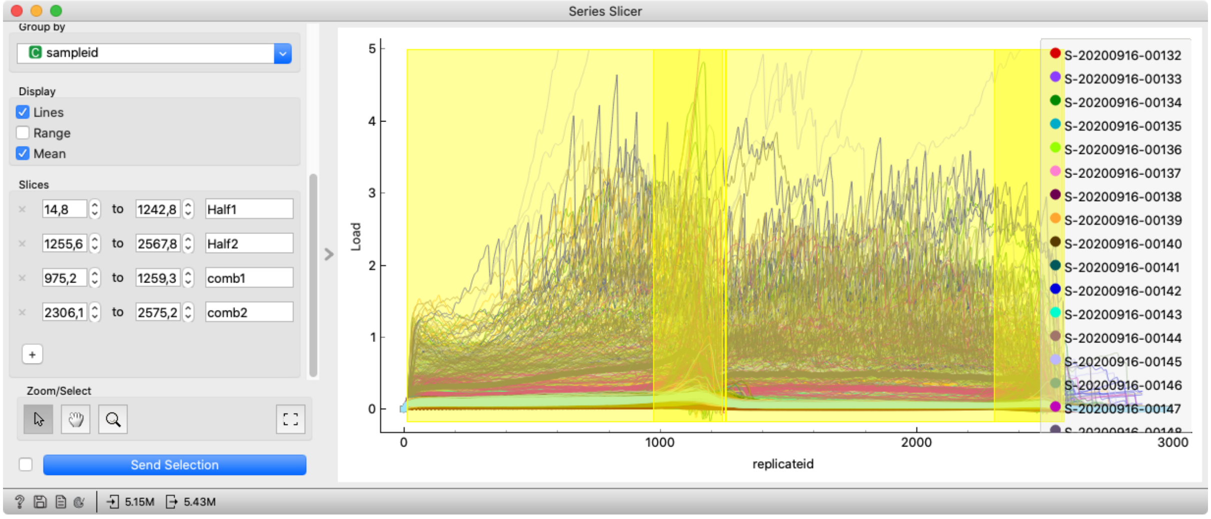
Slices

Zoom/Select

□ Send Selection

235k 235k

We use another *Series slicer* widget on all data, where we select different slices in the data for the aggregation.


Series Slicer

Group by: sampleid

Display
☑ Lines
□ Range
☑ Mean

Slices
14,8 to 1242,8 Half1
1255,6 to 2567,8 Half2
975,2 to 1259,3 comb1
2306,1 to 2575,2 comb2

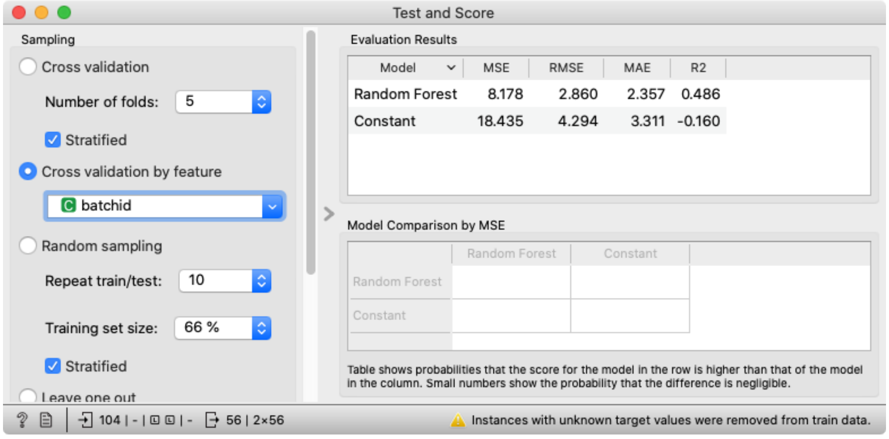
Zoom/Select

□ Send Selection

In the *Aggregate* widget that follows we select `sampleid` and `Tress ID` as Rows, step as Columns, and `Load` as Value to aggregate. We use mean and max aggregations. We use *Merge Data* to add `batchid` and `productdesc` columns to the data.

In the bottom branch we use Select Columns to select "After Use - Rating While Hair Damp Outside of Shower - Easy to Comb (Detangle) Hair When Wet" and "After Use - Rating While Hair Dry and Styled - Easy to Run Fingers/comb Thru My Hair" as attributes and add `BatchID` attribute to meta part of the table. We then average both attributes on the batch level (`BatchID` is set as Rows, both attributes as Values, and Mean as the aggregation). We merge Consumer study aggregations to Intron data with *Merge Data (1)* widget via `BatchID`.

In the last part of the workflow, we use Select Columns to switch between both selected consumer study attributes as a target variable. In *Test and Score* we see that Random Forest with a Mean absolute error (MAE) of 2.357 predict the target variable "After Use - Rating While Hair Damp Outside of Shower - Easy to Comb (Detangle)".



○ Cross validation

Number of folds: 5

☑ Stratified

● Cross validation by feature

batchid

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

Model	MSE	RMSE	MAE	R2
Random Forest	8.178	2.860	2.357	0.486
Constant	18.435	4.294	3.311	-0.160

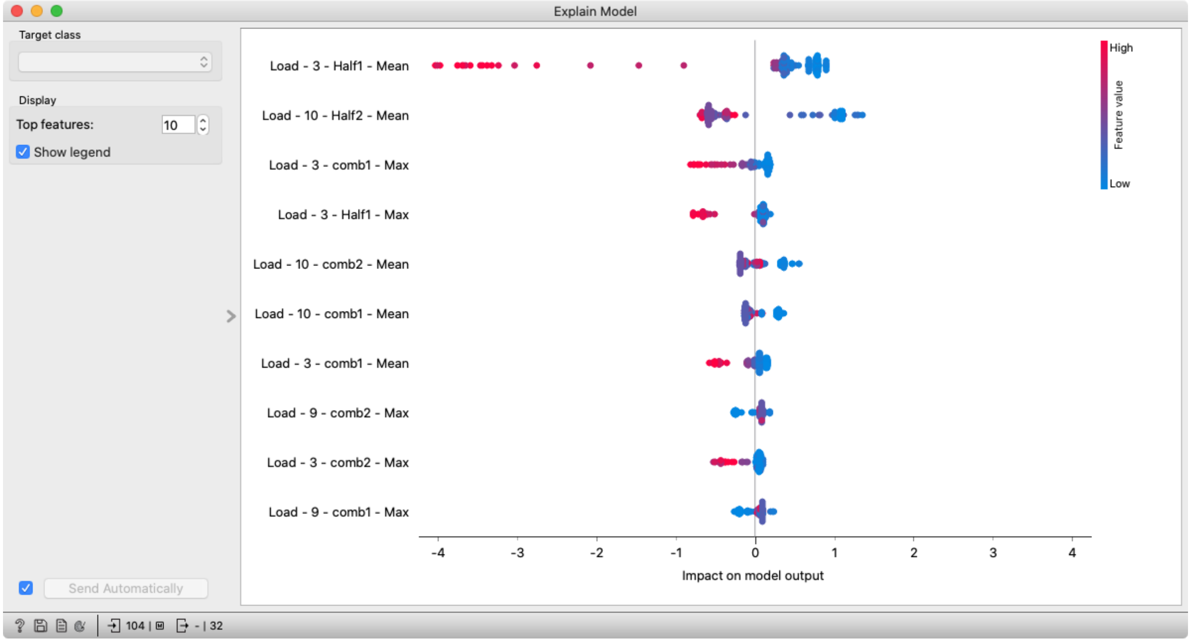
	Random Forest	Constant
Random Forest		
Constant		

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

104 | - | ⊞ ⊞ | - 56 | 2×56

⚠ Instances with unknown target values were removed from train data.

Finally, we will open the *Explain Model* widget which reveals that the mean load from step 3 of the slice Half1 of Instron aggregation is the most important attribute for predicting the target attribute.



Top features: 10

☑ Show legend

We could also try to predict the other selected attributes from the consumer study — “After Use - Rating While Hair Dry and Styled - Easy to Run Fingers/comb Thru My Hair”. What is MAE in this case and does the explanation change and how?

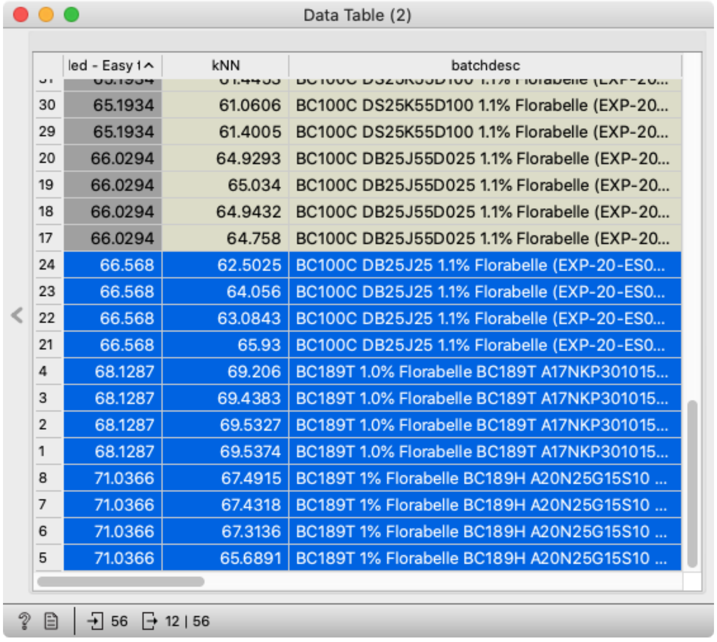
Lesson 3: IFF

We have already performed some explorative analysis of IFF data and seen that the area under the curve and the absolute area under the curve profile the data well. Now we will use that profiling to predict consumer ratings.

Start as before and select a single step with *Select Rows* to simplify the task. After that, use *Aggregate* to profile the data for each `sampleid` and `dataset` combination. Merge in the batch information from Metadata and batch-aggregated responses from the Consumer study. You can try to predict the responses to "Easy to run fingers/comb thru my hair" and "Leaving my hair feeling soft" or find another question you think to be related to IFF measurements.

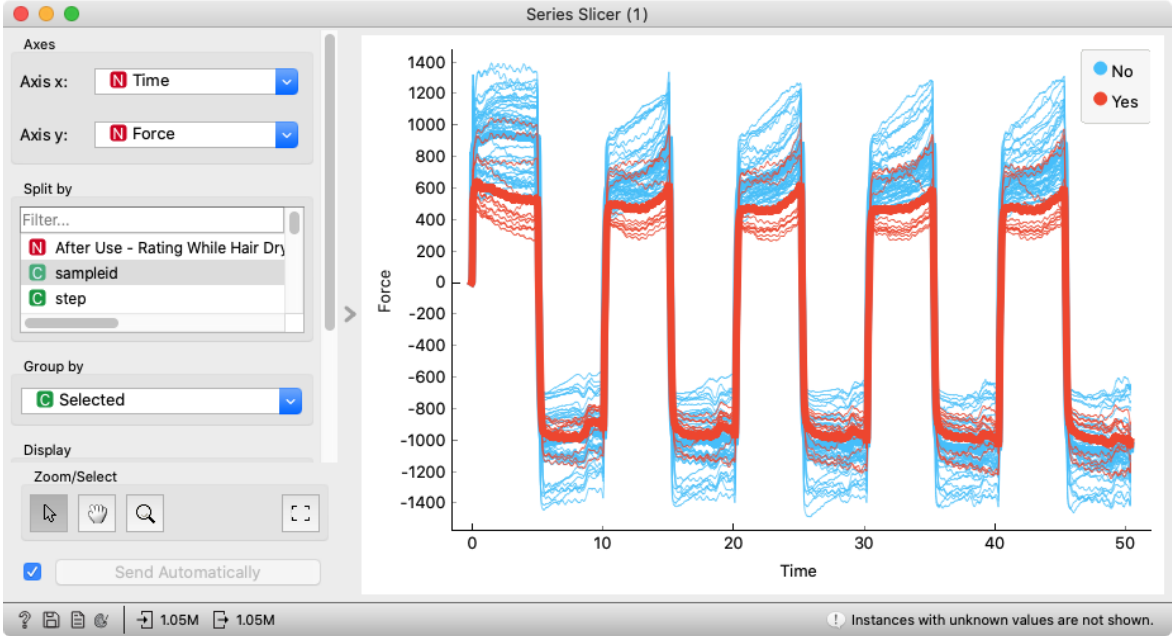
Because we have just 2 independent features, you might want to try some non-linear models in addition to Linear regression (e.g. *Random Forest*, *kNN*). Use *Test and Score* to evaluate different models and choose Cross validation by feature `batchid`.

Test and Score can output predictions. Visualize them in *Scatter Plot* and select the predicted and actual values for axes x and y. You can review the predictions in the *Data Table* as well. As a bonus try to select some rows of interest and merge the Data output (containing the Selected variable) to the raw data and visualize the selection.



Data Table (2)

	led - Easy t	kNN	batchdesc
30	65.1934	61.0606	BC100C DS25K55D100 1.1% Florabelle (EXP-20...
29	65.1934	61.4005	BC100C DS25K55D100 1.1% Florabelle (EXP-20...
20	66.0294	64.9293	BC100C DB25J55D025 1.1% Florabelle (EXP-20...
19	66.0294	65.034	BC100C DB25J55D025 1.1% Florabelle (EXP-20...
18	66.0294	64.9432	BC100C DB25J55D025 1.1% Florabelle (EXP-20...
17	66.0294	64.758	BC100C DB25J55D025 1.1% Florabelle (EXP-20...
24	66.568	62.5025	BC100C DB25J25 1.1% Florabelle (EXP-20-ES0...
23	66.568	64.056	BC100C DB25J25 1.1% Florabelle (EXP-20-ES0...
22	66.568	63.0843	BC100C DB25J25 1.1% Florabelle (EXP-20-ES0...
21	66.568	65.93	BC100C DB25J25 1.1% Florabelle (EXP-20-ES0...
4	68.1287	69.206	BC189T 1.0% Florabelle BC189T A17NKP301015...
3	68.1287	69.4383	BC189T 1.0% Florabelle BC189T A17NKP301015...
2	68.1287	69.5327	BC189T 1.0% Florabelle BC189T A17NKP301015...
1	68.1287	69.5374	BC189T 1.0% Florabelle BC189T A17NKP301015...
8	71.0366	67.4915	BC189T 1% Florabelle BC189H A20N25G15S10 ...
7	71.0366	67.4318	BC189T 1% Florabelle BC189H A20N25G15S10 ...
6	71.0366	67.3136	BC189T 1% Florabelle BC189H A20N25G15S10 ...
5	71.0366	65.6891	BC189T 1% Florabelle BC189H A20N25G15S10 ...



Series Slicer (1)

Axes

Axis x: N Time

Axis y: N Force

Split by

Filter...

N After Use - Rating While Hair Dry

C sampleid

C step

Group by

C Selected

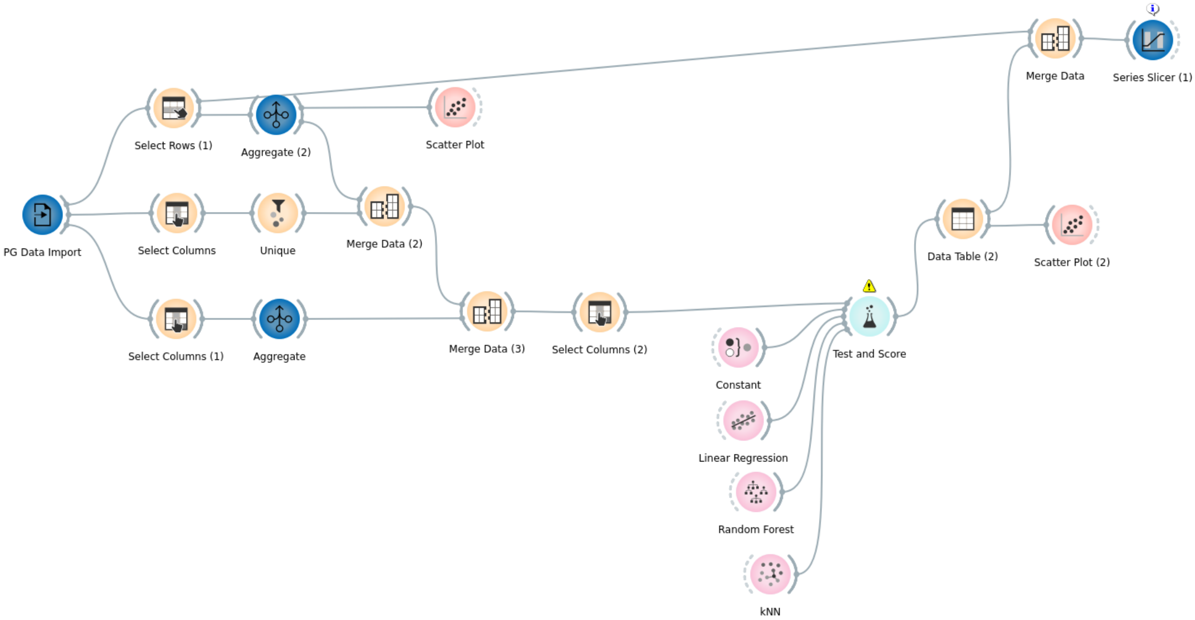
Display

Zoom/Select

☑ Send Automatically

1.05M 1.05M

! Instances with unknown values are not shown.



Merge Data
Series Slicer (1)
Select Rows (1)
Aggregate (2)
Scatter Plot
PG Data Import
Select Columns
Unique
Merge Data (2)
Data Table (2)
Scatter Plot (2)
Select Columns (1)
Aggregate
Merge Data (3)
Select Columns (2)
Test and Score
Constant
Linear Regression
Random Forest
kNN