Drivendata Cleaning time

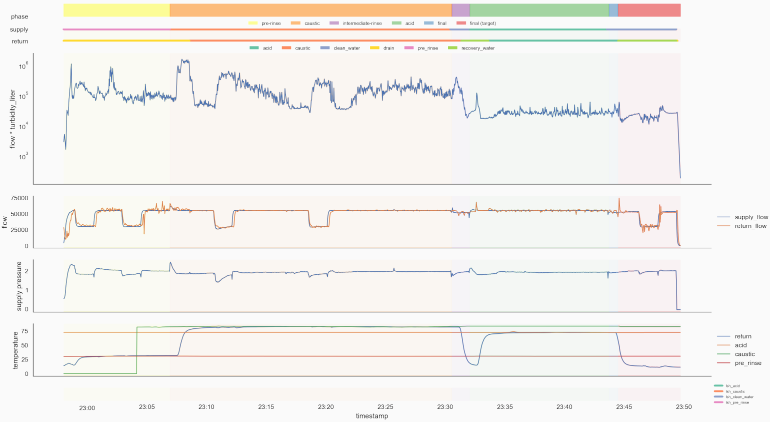
Stage 2: Modeling Report Competition

Tomasz Waleń

Dataset - overview

- dataset contains data about 7988 cleaning processes
- each process is described with time-series (2 sec freq) with:
  - measured turbidity
  - temperature readings
  - supply pressure
  - flows
  - tank levels
  - type of activity on the supply / return line (one-hot encoded)
  - low levels flags
- train data consisted data from 2018-02-21 until 2018-04-25, test data 2018-04-25 until 2018-05-24

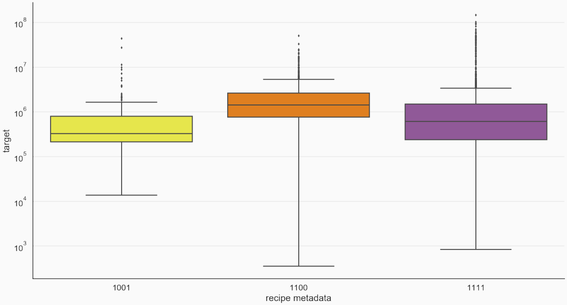
How does the sample process look like?



Target phase is marked with the red color.

Most important features - recipe metadata

Recipe meta data is not enough to guess target value, but still give some information about the range and standard deviation of the targets.

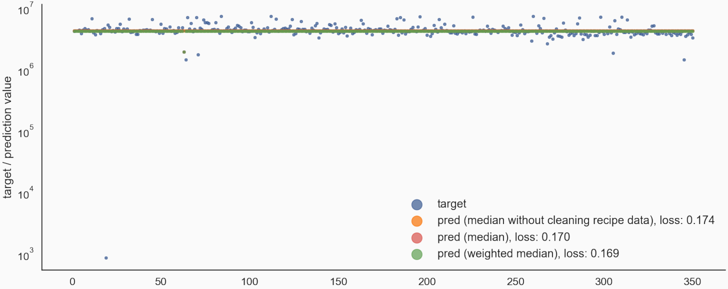


For example process without intermediate-rinse and acid phases tend to have higher target values (and very high std. dev.).

Most important features - target means

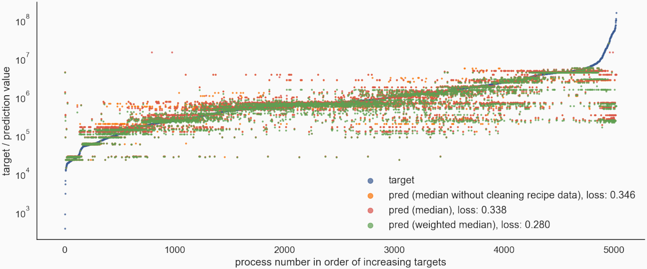
For single object most target values do not differ, so quite good strategy is to predict using mean target value.

Let's look at the target values for object 405:



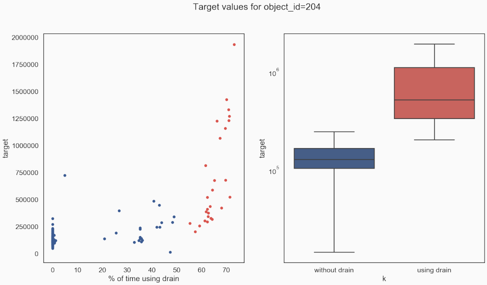
Most important features - target means continued

We can enhance procedure even more using weighted median instead of mean (weight=1/target). During contest the dataset was enhanced with recipe data. If the target mean is computed with weighted median for each combination of (object_id, recipe) this method alone gives 0.280 loss. This creates huge data leak so it was not so effective during the contest, but still it was possible to score 0.320 with this method alone.



High target value processes

I've tried to understand why some processes end up with very high target values. I've noticed that, in the target phase, the majority of the processes return water into pre-rinse tank, but some dump it into the drain (which is typical only for pre-rinse phase). It turned out that such process are also very likely to have high turbidity.



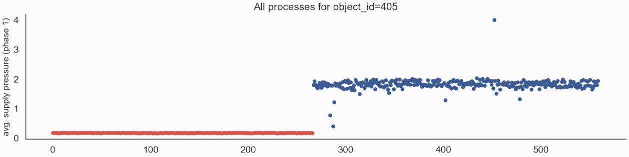
Sometimes such behaviour could be detected in earlier phases.

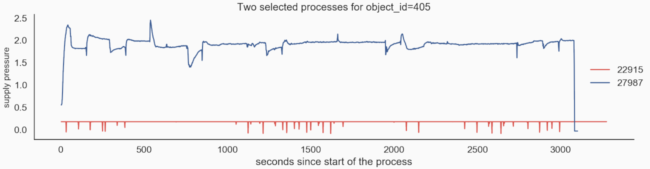
Data phenomenons - unreliable pressure data

In the training dataset some pressure readings were unreliable (i.e. for object 405).

Before 2018-04-12 all `supply_pressure` values were limited to 0.17.

After 2018-04-12 the mean `supply_pressure` was around 2.0.





Data phenomenons - negative turbidity

Surprisingly the `return_turbidity` could be negative:

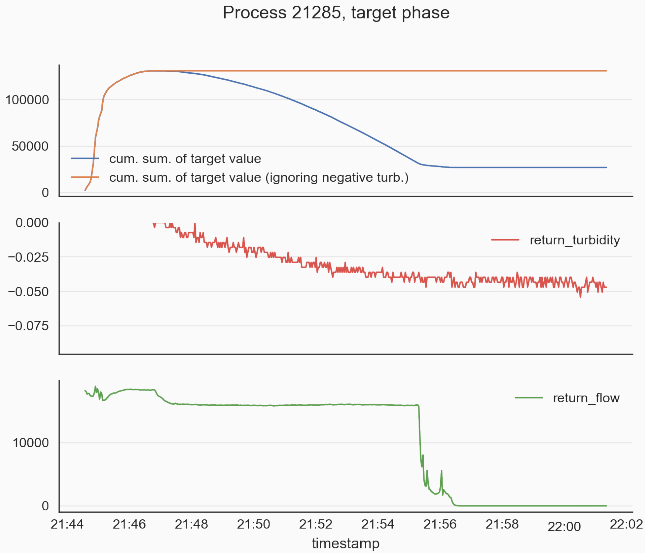
- nearly 3.9% of processes contain entries with negative values
- nearly 1.5% of processes contain entries with negative values in target period

The formula for calculating target value takes into account negative flow, but does not takes into account negative turbidity, and in such case it could make the difference. In some cases ignoring negative turbidity could change the target value by 400%.

I was not able to understand meaning of negative entries, also it was hard to predict in advance whatever given series would contain such effect.

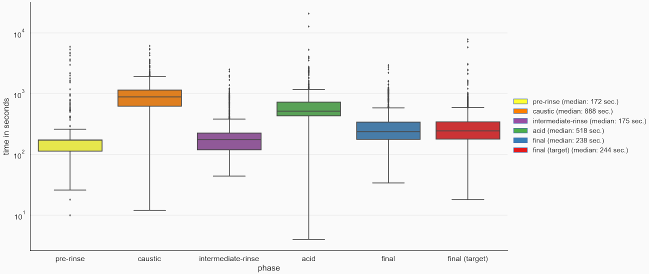
Data phenomenons - negative turbidity an example

For example such behaviour could be seen in the following process:



Data phenomenons - duration of phases

There is large variety in the duration of the processes & and the duration of the processes.



Some processes are clear outliers and possible result of an error:

- process 22112 with the acid phase taking almost 6h
- process 26113 with the acid phase taking 4 sec.