Rinse over run

Predictions and insights for a more sustainable industry

Executive summary

Model, created to predict the final turbidity of the cleaning process, has a MAPE score on the test set of 0.2747. The most predictive input for the model is the object that will be cleaned up. This suggests that object metadata like the material, the geometry and the kind of dirt are the most important predictors for the model

After a definition of the most important factors we focused on the process-configurable variables like duration, flow and pressure. These features don't have a strong predictive power and they don't give big improvements to the model. This does not imply that is not important to tune them, instead it highlights that the process configurations are standardized leading to low space for statistical learning

After a deep analysis we have identified two main ways to improve the model. If the business goal is to predict the outcome of the cleaning process adding object metadata will lead to the best results. If the business goal is optimize the processes and reduce costs it's necessary to add new data with non standard process configuration

The turbidity prediction journey: 4 steps from data preparation to insights generation

1 Data preparation and feat engineering

- New features creation to improve physical description:
  - Sum of flows total liters used in the process
  - Multiplication between valve state and flows
  - Difference between supply and return flow
- More than 200 features have been tested

2 Model training and validation

- Model build on time series data 5M observations
- Create 2 layers of stacked models with lightGBM (14 models for layer 1, and 1 model for layer 2)
- Validation set created with 3-folds time stratified split

3 Aggregate predictions

- Aggregate predictions of second layer model with a weighted median (aggregated 1.880.160 test prediction to produce the final 2967ones)
- Validation score consistent with leaderboard: 0.270 for the best model

4 Analysis and insights

- Investigate the model outputs to understand what information is more valuable and how to improve the predictive power

Focus of this presentation

Analysis: feature importance, model confidence and feature distributions deep-dive reveals main correlations

Feature importance

a

Goal: Understand where is stored the most valuable information

Standard approach to do it:
Use lightGBM built-in methods

Too much noise, not stable for this use case

Solution:
Build several monovariate models and compare the single feature performance with the best constant prediction

Model confidence

b

Goal: Understand when prediction are more reliable

Standard approach to do it:
Take the validation predictions, build a model that try to predict the error

LigthGBM builds on residuals so the results don't have a lot of sense

Solution: Observe how the confidence on the validation set change against each feature

Feature distributions

c

Goal: Understand how the model learn and spot possible data issues

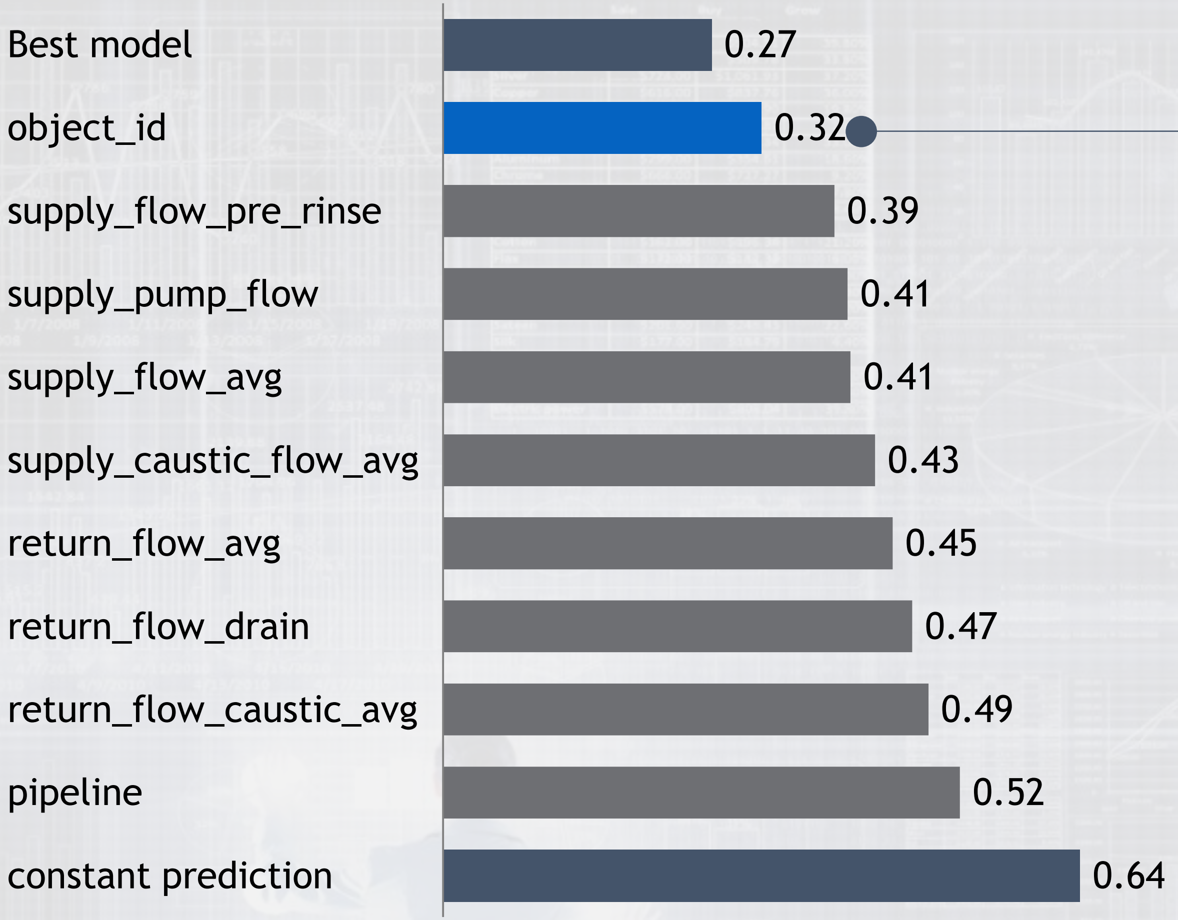
Standard approach to do it:
Observe the distribution of each feature

Tree based model split on the strongest features first. So the distribution of the less important features are sampled from the first one

Solution:
Split on the strongest feature and study other features distributions

a Feature importance: object_id the most important for prediction

MAPE score of most performant monovariate models



It's possible to achieve a top 30 in the leaderboard just with object_id

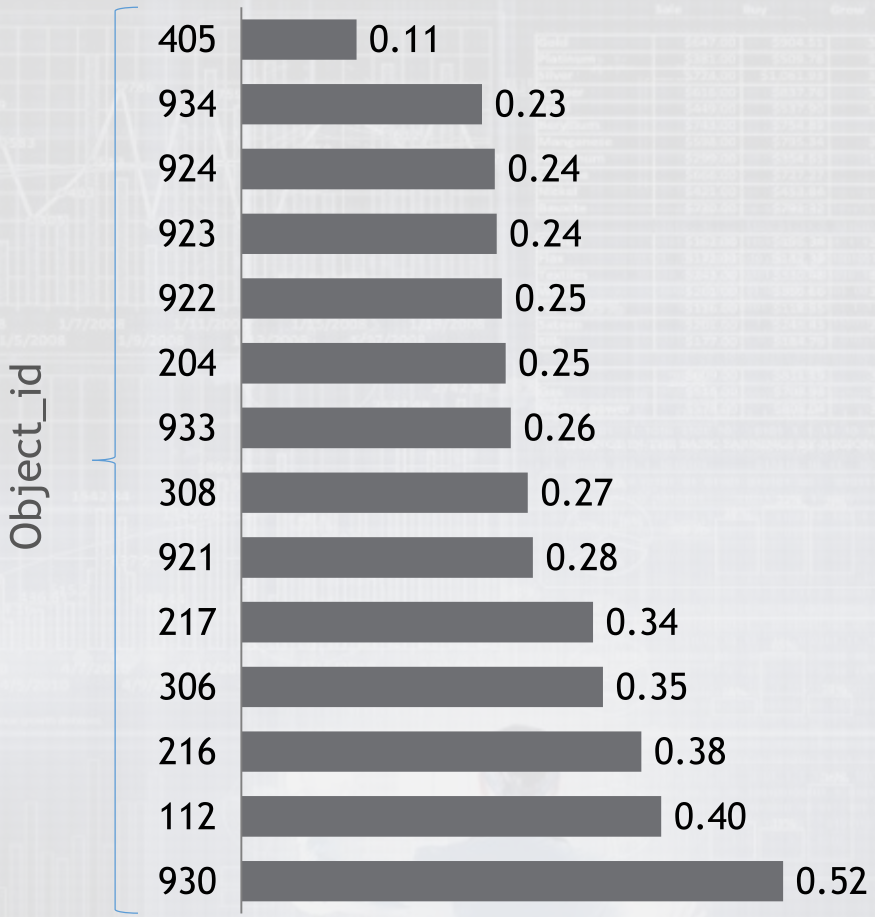
Adding other features leads to a model that score 0.27 (42% of the constant prediction); a relatively small improvement compared with the 50% achievable with only object_id

Which object is cleaned (object_id) is the best insight on the final outcome of the process; this means that the **material, the geometry and the kind of dirt on the object play a fundamental role**

NOTE: Constant prediction model outputs always 291 000 liters; All the score refers to the validations set for consistency; The list include engineered features not present in the original dataset

b Model confidence: the object impacts the process predictability

Best model MAPE score by object id



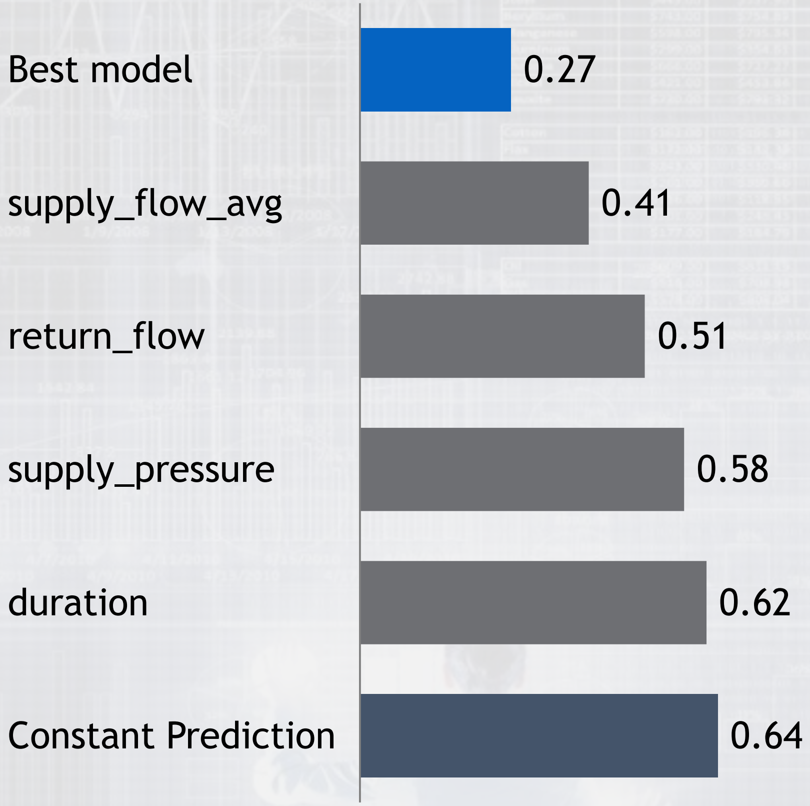
- Different objects show a different confidence level of the outcome
- This suggest that the same object can be in different initial states

- A set of variables for describing the pre-rinse object state can reduce the unpredictably of some objects; e.g. :
    - How much the object was dirty?
    - What was last cleaning time?
    - How many operative hours since last cleaning?

NOTE: All the score refers to the validations set for consistency; Only a short list of the Object id is presented

a Feature importance: process configurable variables lack in predictive power

MAPE score of process optimizable features



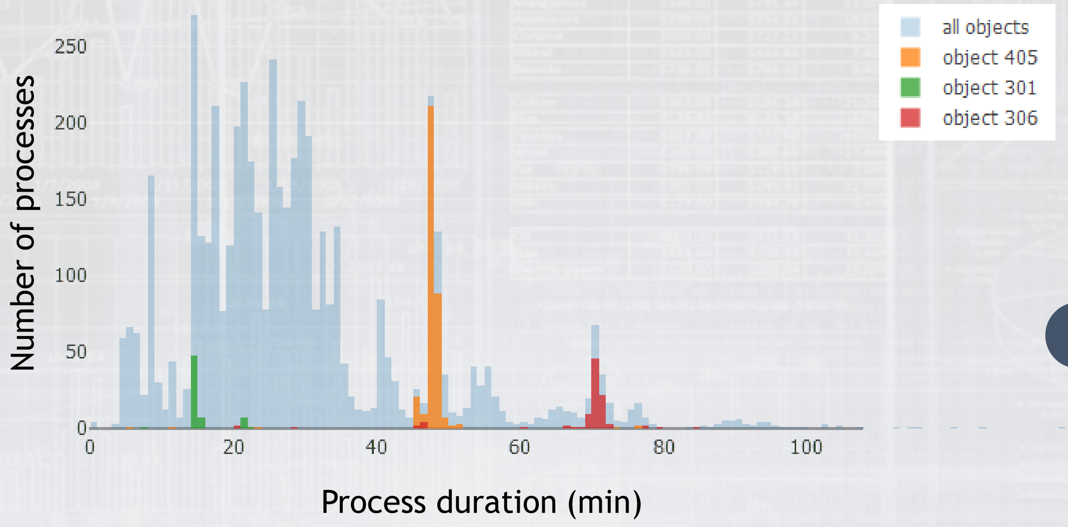
The process configurable variables, that are modifiable from the operators, don't seem to impact much the final outcome of the process.

- In the dataset there are standardized processes for each object
- This reduce what the model can learn from the configurable features, resulting in apparently irrelevant actionable variables

Details next slide

c Feature distribution: standardized processes lead to less space for learning

Process duration by object id



Process duration distribution is very narrow and object dependent; **this impacts the predictive performance** (the same happens for flows and pressures)

Narrow distributions means that the model has not seen how the target variable changes with the covariate (e.g.: washing t-shirts always a 30 °C does not provide insights on what is going to happen at 90 °C)

Increment variability in the cleaning process - this will increase the information in the data and will make the model useful for process optimization

NOTE: The distribution in the histogram is created from the whole train set, final rinse phase excluded

Insight: 2 ways to improve model and processes

a+b Add object metadata

Consider to add in the model:

- Geometry
- Material composition
- Dirty composition
- Initial state

This will improve the model predictive power

a+c Test non standard process parameters

Consider testing different cleaning configurations for each object:

- Durations
- Flows
- Pressures

This will increase model sensibility to process configurable variables, leading to a more optimizable and cost effective processes

Details next slide

Personalize the process for each object and its state

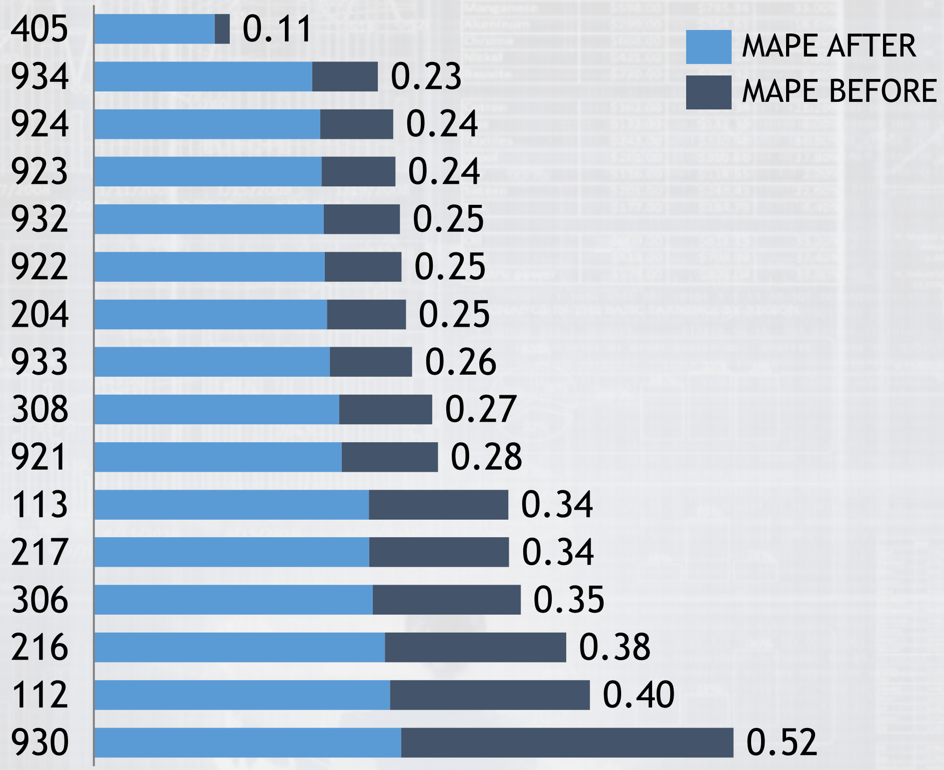
Include process metadata and non standard processes to push optimization even further - explore best solution in an algorithmic way and find best cleaning solution depending on the object and its cleaning state

Insight: adding object metadata and standard processes will improve model MAPE

Illustrative – qualitative indications

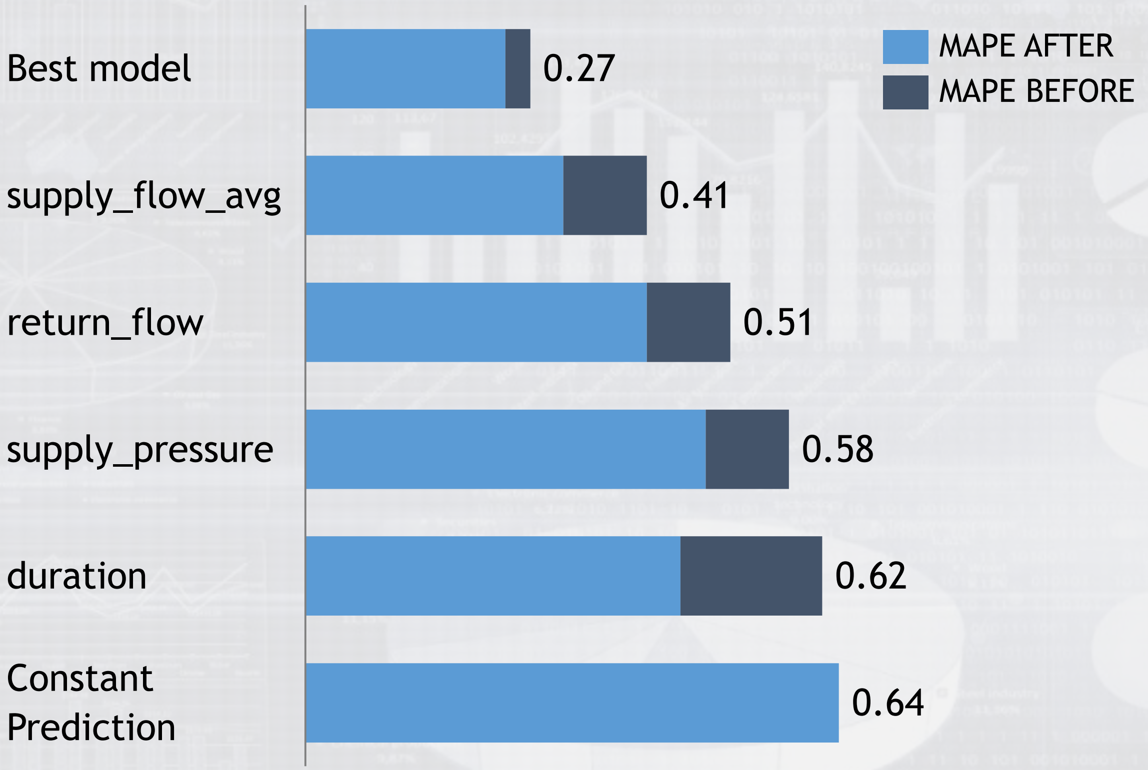
a Add object metadata

MAPE improvement by object id



b Try non standard processes

MAPE score improvement for process configuration



NOTE: the score improvements are hypothetical and must be checked with data. They serve only as illustrative scope