DRIVENDATA
Sustainable Industry: Rinse Over Run

Stage 2: Modeling Report Competition

Name : **Bernd Allmendinger**
Username: **neurocomputing**

Start Situation

- **Stage1: The model with the best cv and the best private leaderboard score is a single gradient boost model with public score = 0.2719, private score = 0.2765 and 5 fold cv score = 0.2672.**

- Working with a single model has multiple benefits over more complex systems:
  -straightforward variable importance assessment
  - ease of prediction of confidence level
  - simple sensitivity analysis

- The model was fitted with *LightGBM*. The number of iterations and features selection was determined by a 5 fold cross validation with the build in function `lgb.cv`.

- The optimization of the hyper parameters and features selection was done by a 5-fold cross validation.

- **Structure of feature name**:
  - The letter *' P '* followed by a number indicates the phase (e.g.'P2' refers to the statistics collected from phase 2)

  - The four phases are coded as follows:
    - *pre_rinse* = 1
    - *caustic* = 2
    - *intermediate_rinse* = 3
    - *acid* = 4

  - If a feature names ends with *' Head5 '*, the statistics of the first 5 observations are used as a feature (`lambda x: x.head(5).mean())`
  - If a feature names ends with *' Tail5 '*, the statistics of the last 5 observations are used as a feature (`lambda x: x.tail(5).mean())`

  - If after the phase the string *' DiffTailHead*' follows, the statistics of the last 5 observations of the phase is subtracted with the statistics of the first 5 observations
  - If the feature starts with the string *'Phasediff*' we subtract the statistics from two consecutively phase

Sensitivity analysis – Method 1: *LightGBM* feature importance

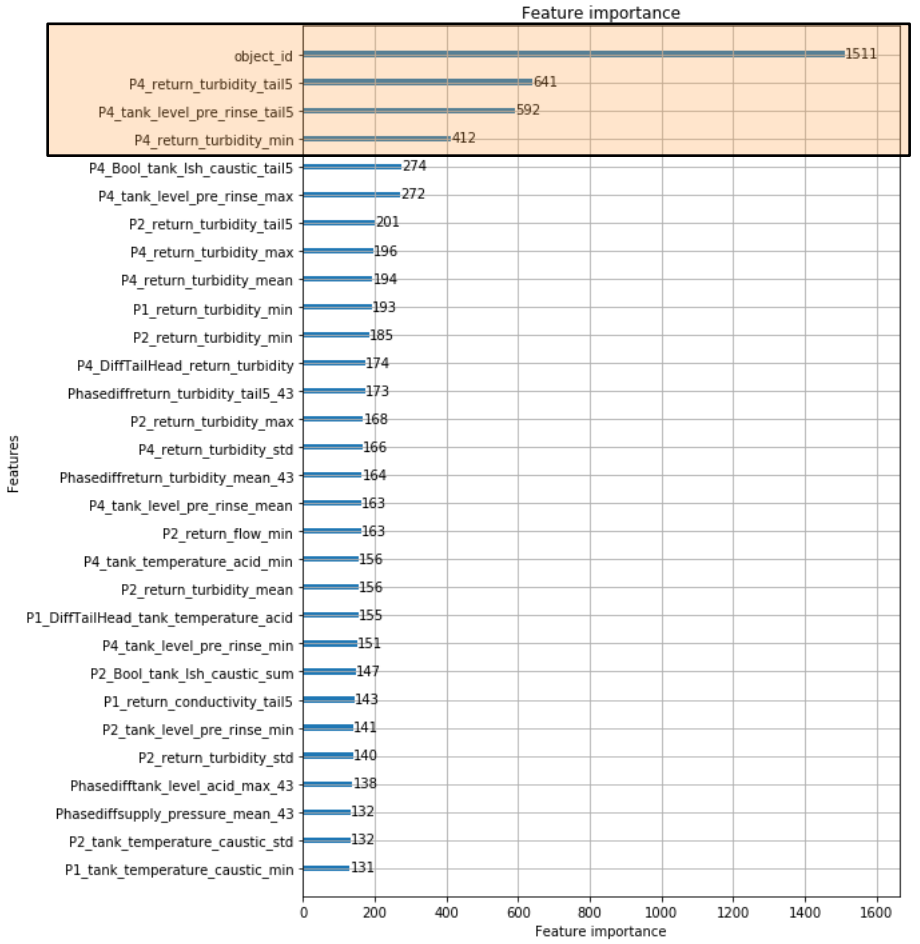
LightGBM has a function (`lgb.plot_importance`) which calculates the feature importance from the derived features.

The disadvantage is that the influence of the raw variables on **rinsing phase outcome** can not be determined directly, because many derived features from the time series are calculated from the raw variables.

The Top 4 derived features with the highest impact on the rinsing phase outcome are:

1. *Object_id*
2. Mean of last 5 values of *return_turbidity* in phase *acid*
3. Mean of last 5 values of *tank_level_pre_rinse* in phase 4
4. Minimum value of *return_turbidity* during phase *acid*

The fact that the *object_id* has such a large influence shows that every factory has highly specific features and the CIP strongly depends on the installed equipment characteristics and recipe/product portfolio.

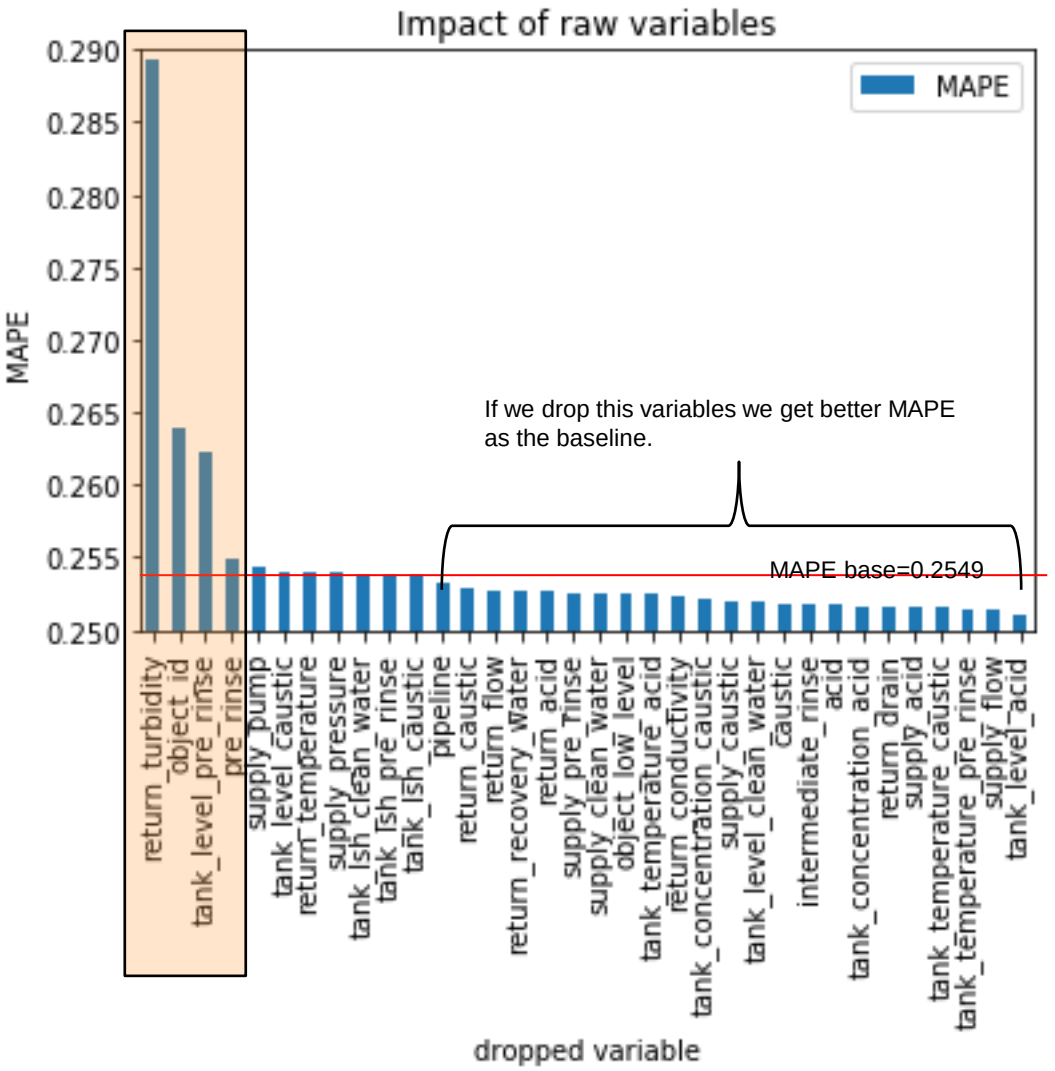


Sensitivity analysis– Method 2: Variable drop out

"**Variable drop out**": For each "raw" variable, a model has been created and fitted excluding this specific variable.
The MAPE metric from a 5 fold cross validation is used to measure the influence of the raw variable.

The increment of the cross validation MAPE metric for the excluded variable gives a real indication of the influence on the rinsing phase outcome.

As a base line we use the cross validation MAPE score from the model fitted wit all 35 variables.
MAPE base = 0.2549



The four variables with the highest impact on the rinsing phase outcome are the following:

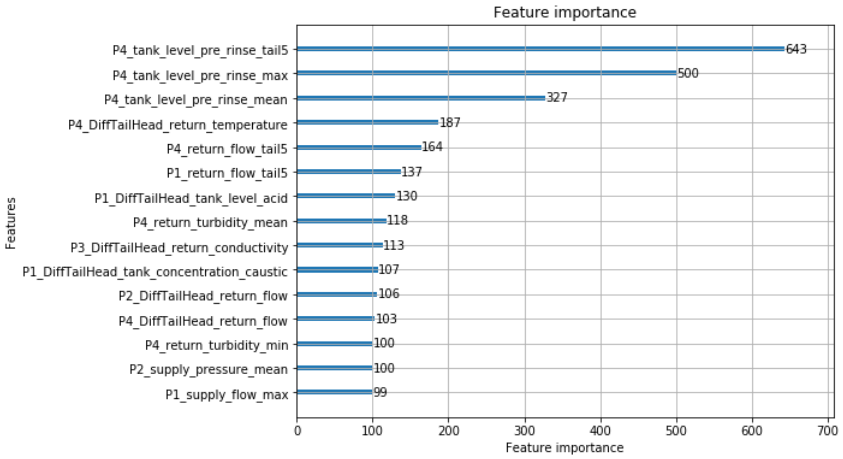
1. *return_turbidity*
2. *Object_id*
3. *tank_level_pre_rinse*
4. *pre_rinse*

A similar conclusion can be drawn using the two methods.

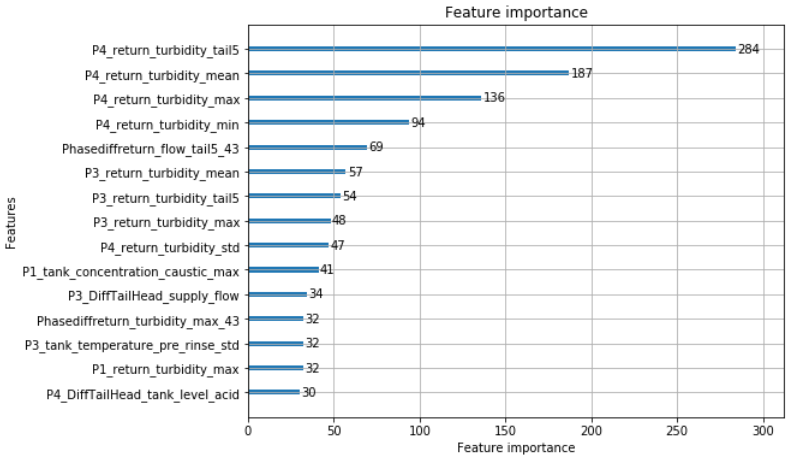
Sensitivity analysis – Object specific

In the last two analysis we have seen that the influence of the variables on the rinsing phase outcome depends strongly on the object.
In order to derive valuable information for the daily work, it is important to perform an object-specific analysis.
To illustrate the differences between the objects, the analysis is performed on the two objects with the most process ids (405 and 932). We train a *LightGBM* grading boost model for each object. To prevent overfitting, we determine the number of iterations by a 5 fold cross validation

Object 405 **cv-MAPE: 0.1027**



Object 932 **cv-MAPE: 0.2023**



The two charts show the Top 15 most important features for object 404 and 932.
The features with the greatest influence on the calculation of rinsing phase outcome are different in both objects.
For the object 405, the max, min and tail signals for the variable *tank_level_pre_rinse* in phase *acid* have the biggest influence.
While for the object 932, the signals of variable *return_turbidity* in phase *acid* have the biggest impact.
The difference is statistically significant because the correlation between the two raw time series is low (0.006).

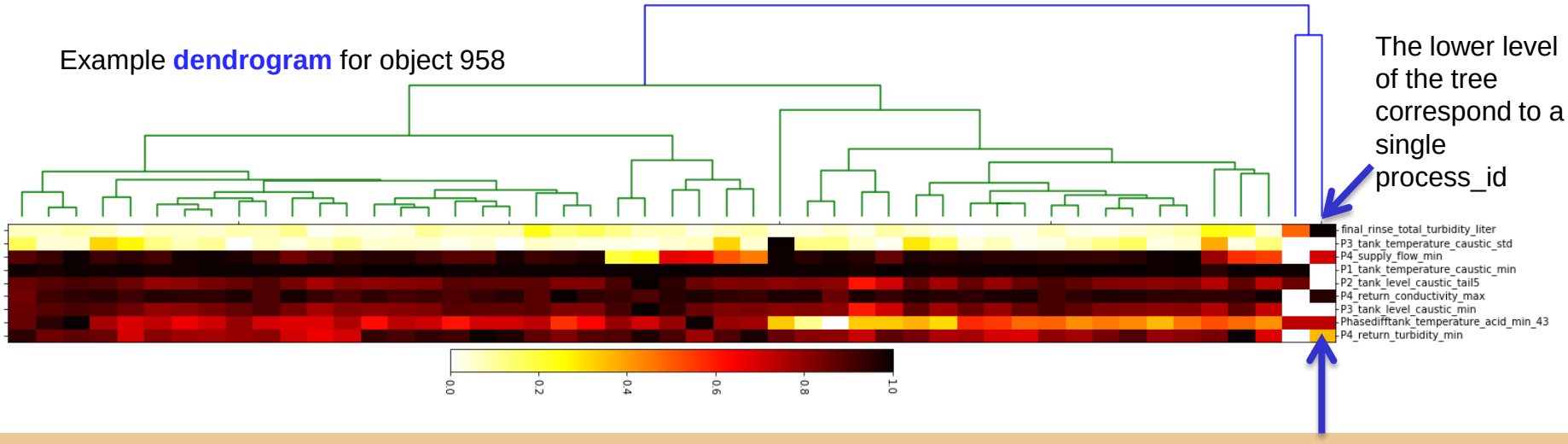
Clustering of physical world situations/events

Which signals at which moment are mainly responsible for the presence of turbidity during the final rinse ?

To answer this question a **agglomerative hierarchical cluster** procedure is applied to the data.

1. The complete linkage procedure from the scipy-package is used as an algorithm for the agglomerative hierarchical cluster procedure.
2. To better interpret the results, the clusters are visualized with a **dendrogram**. To link the clusters to the features values, a heat map of the features is added to the **dendrogram**.
3. Since the previous sensitivity analysis has shown that the local conditions in the object are imported, the analysis is performed per object.
4. The top 8 feature with the most influence from the previous sensitivity analysis, plus the rinsing phase outcome is used for clustering.
5. The features are linearly transformed via *MinMaxScaler* from *scikit-learn* into the interval [0,1].

Example **dendrogram** for object 958



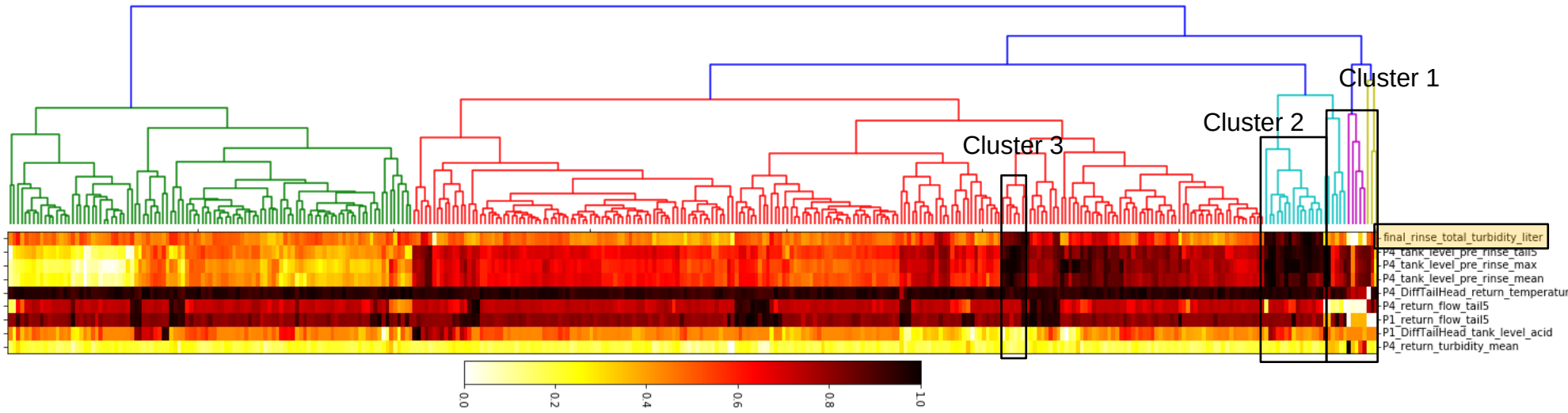
The lower level of the tree correspond to a single process_id

The first process_id in the dendrogram has a very high *turbidity during the final rinse* (black color) , the features *P3_tank_temperature_caustic_std*, *P1_tank_temperature_caustic_min*, *P2_tank_level_caustic_tail5*, and *P3_tank_level_caustic_min* are all at their lowest value (white color).

Clustering of physical world situations/events – Object 405

Which signals are responsible for the presence of turbidity during the final rinse for object 405?

In the **dendrogram** at least 3 clusters are shown.



Clusters properties:

Cluster 1: All *process_id's* in this cluster have low *turbidity during the final rinse* (bright color), low *return_flow* at end of phase *1=pre_rinse* and 2. The minimum, maximum and tail signals of *tank_level_pre_rinse* at the end of phase *4=acid* are around 80%.

Cluster 2: In this cluster all *process_is's* have a high (about 90%) *turbidity during the final rinse* (dark color). All signals/events (max, mean and tail) from *tank_level_pre_rinse* in phase 4=acid have very high values of about 90%. The signals (tail-value) from *return_flow* have medium values (about 60%) at end of phase 1=*pre_rinse* and 2=*acid*.

Cluster 3: This cluster is very similar to Cluster 2 with very high values for *turbidity during the final rinse*. The difference lies in the signal *P4_DiffTailHead_return_temperature* which is lower in this cluster than in cluster 2

Level of confidence for predicted outcomes

Quantile regression is used for the gradient boost model to determine the level of confidence for the predicted outcomes.
Three quantile regressions models where fitted using the same data. Alpha is the parameter for the quantile.
The first model is fitted to predict the median (alpha = 0.5). This corresponds to the fit of a model with the object function MAE (Mean Absolut Error).
The second model is fitted with alpha = 0.1 to predict the lower 10%-confidence level
The third model is fitted with alpha = 0.9 to predict the upper 90%-confidence level

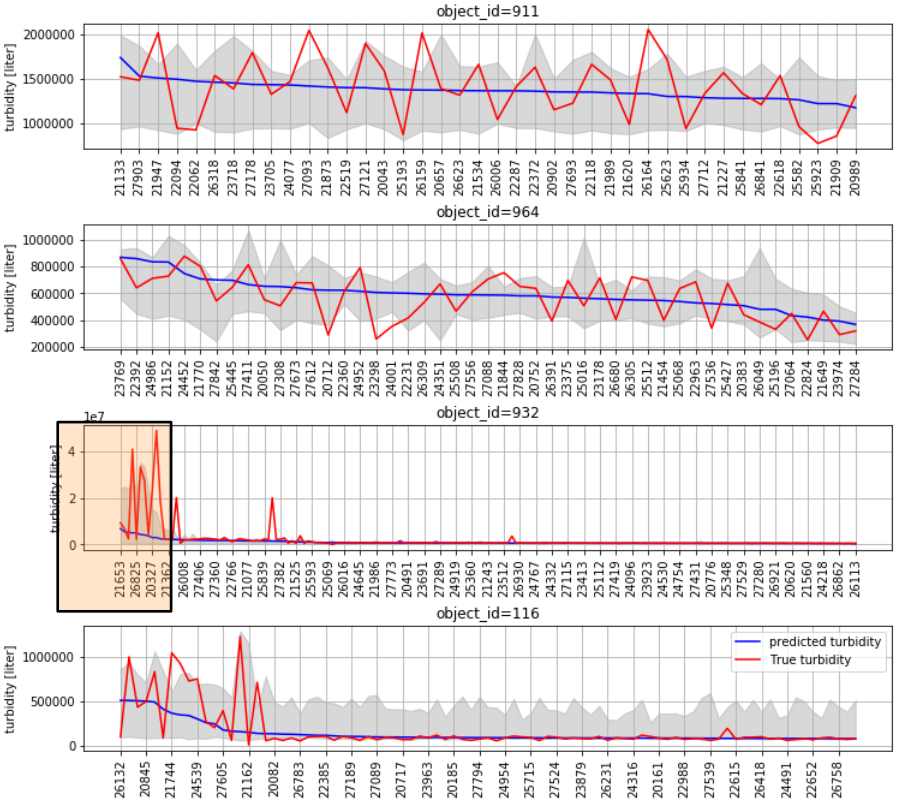
LightGBM offers the possibility to fit the model with object as quantile regression.
To show that the method works, we did a 5 fold cross validation. We train the three models with 5 folds and do the prediction of the median and the lower and upper confidence levels with the observations from the test fold.

The diagram (right side) shows the out-of-fold median predictions (blue), the true values (in red) and the confidence levels for each process_id (x-axis) of for 4 different object_id's.

The lower (=0.1) and upper (=0.9) confidence levels are colored gray.
The out-of-fold predictions were sorted by the predicted values (*final_rinse_total_turbidity_liter)* to better see the influence of confidence levels.

The confidence levels are structurally very different for the various objects.

For object 932 the confidence levels for the first process_id's are large and then becomes thinner.



Level of confidence for predicted outcomes – feature impact on upper confidence level

Which features does impact the upper 0.9-confidence level?

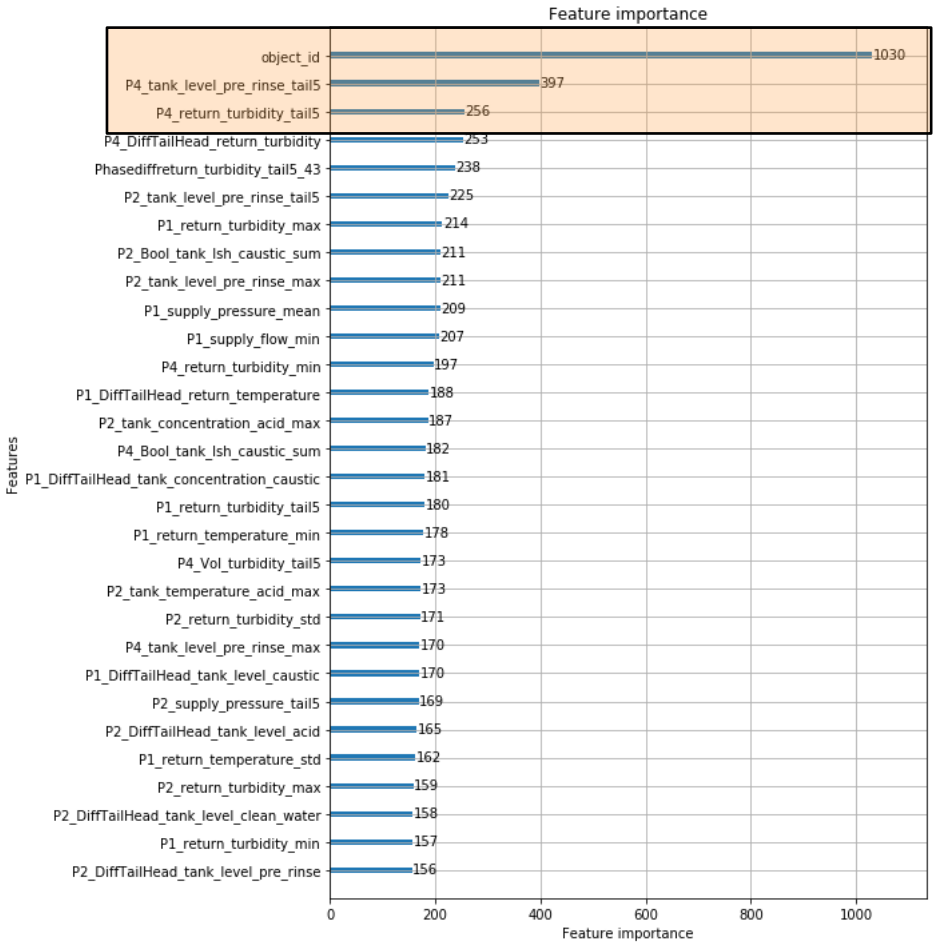
To answer this question we train a *LightGBM* model with all observations.

As objective function we used quantile with alpha=0.9.
To measure the impact of the features, we used the feature importance build in function from *LightGBM.*

From the chart we see, that the top 3 most important features are the same we have for the median model

1. *Object_id*
2. Mean of last 5 values of *tank_level_pre_rinse* in phase *acid*
3. Mean of last 5 values of *return_turbidity* in phase *acid*

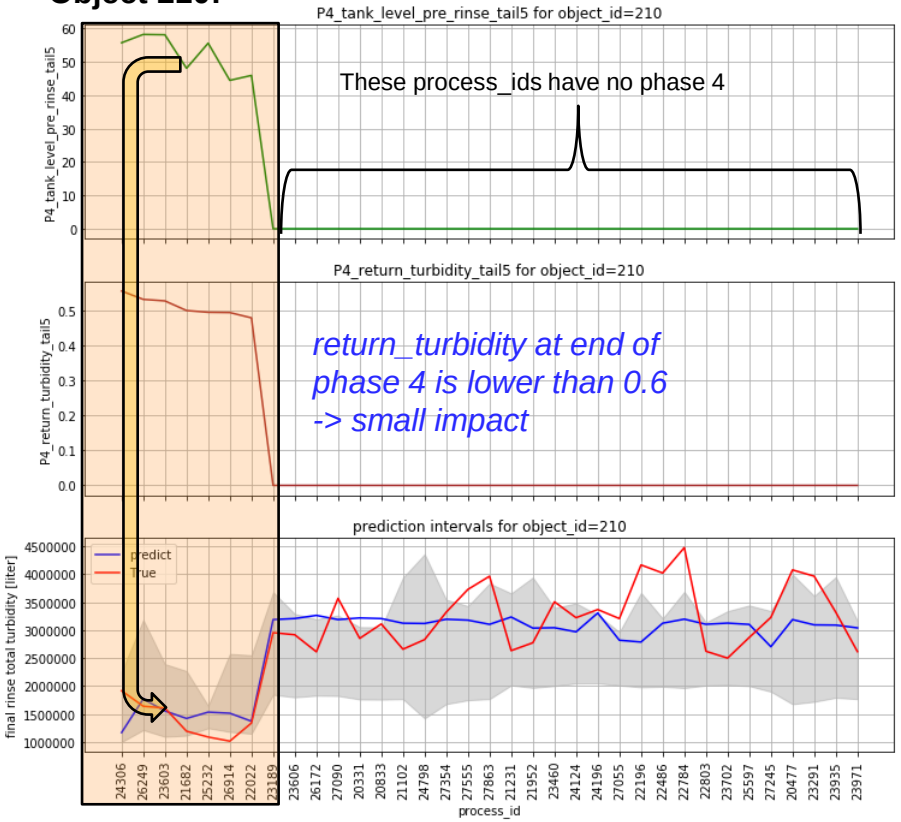
The *object_id* has a very large impact, which also explains that the structure of the confidence intervals between the objects is different



Level of confidence for predicted outcomes – compare with reality

What influence does *return_turbidity* and *tank_level_pre_rinse* have on the rinsing phase outcome?
We check this for objects 932 and 210

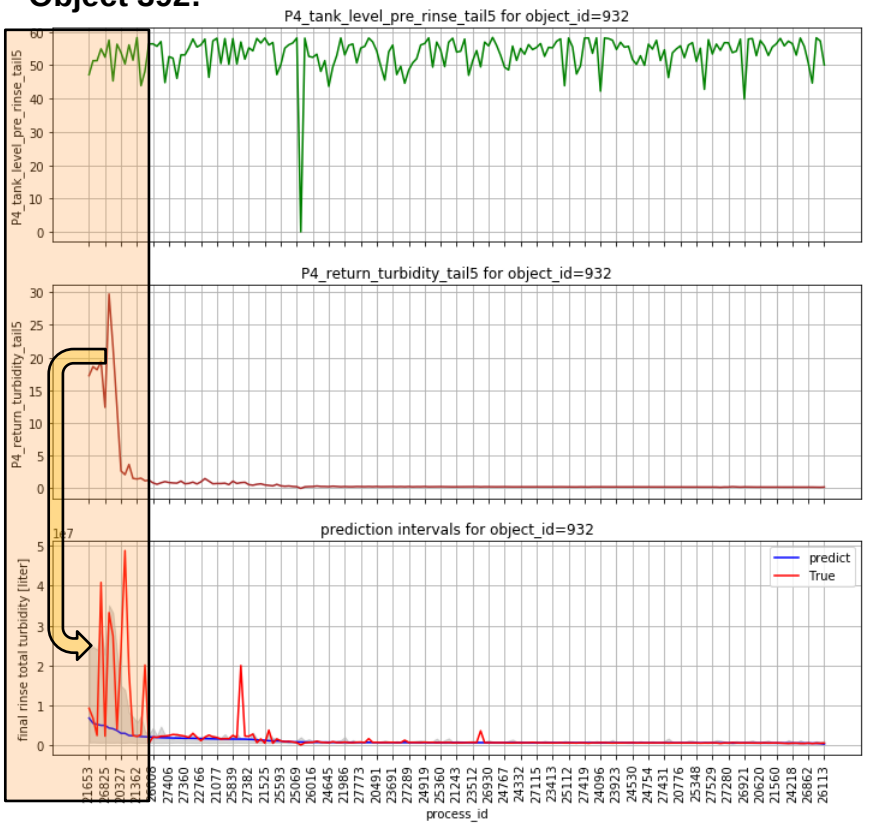
Object 210:



return_turbidity at end of phase 4 is lower than 0.6 and has a small impact on the rinsing phase outcome. (compare with object 392)

High values of *tank_level_pre_rinse* in phase *acid* implies a low rinsing phase outcome.

Object 392:



High values of *return_turbidity* in phase *acid* implies a high rinsing phase outcome.

tank_level_pre_rinse does not seem to have much influence on rinsing phase outcome for object 392.

This example makes it clear, that no general rules can be derived. The behavior depends very strongly on the object.

Summary

- **Stage 1**: The model with the best cv and the best private leaderboard score was a single gradient boost model with object function **MAPE**.
  Lightgbm was used to fit the model. Feature selection and hyper parameter tuning was done by a 5 fold cross validation with the build in function lgb.cv.
  The same model was used for the sensitivity analysis and level of confidence prediction. For the level of confidence prediction, only the object function was changed.

- Two methods are compared for the calculation of the **variable importance**:
  1. with build function for **feature importance calculation** from *Lightgbm* and
  2. "**Variable drop**" method

  The results from both methods are similar.
  The variable importance values are an average value over the objects. To get more meaningful values we fit a *Lightgbm* model for each object. The results show, that each object has different variable (=events/signals) which are important for the calculation of the outcome of the final rinse.

- To find the signals that are responsible for the presence of turbidity during the final rinse a **agglomerative hierarchical cluster** procedure was used. The clusters are visualized with a **dendrogram**. An addition features heat map was linked to the **dendrogram** to recognize the relationships between the signals. The analysis was performed per object.

- The level of confidence for the predicted outcomes was determined by a quantile regression gradient boost procedure. The three models were fitted with *lightgbm*. For the level of confidence determination only the object function was changed to **quantile**.
  The fitted models for the lower (0.1), median (0.5) and the upper level (0.9) are quite stabile.

  To show that the method works, we did a 5 fold cross validation. We train the three models with 5 folds and do the prediction of the median and the lower and upper confidence levels with the observations from the test fold.

  In order to determine the signals responsible for the high level of confidence, a variable importance analysis was performed with the model for the upper (0.9) confidence level.