Rinse over Run

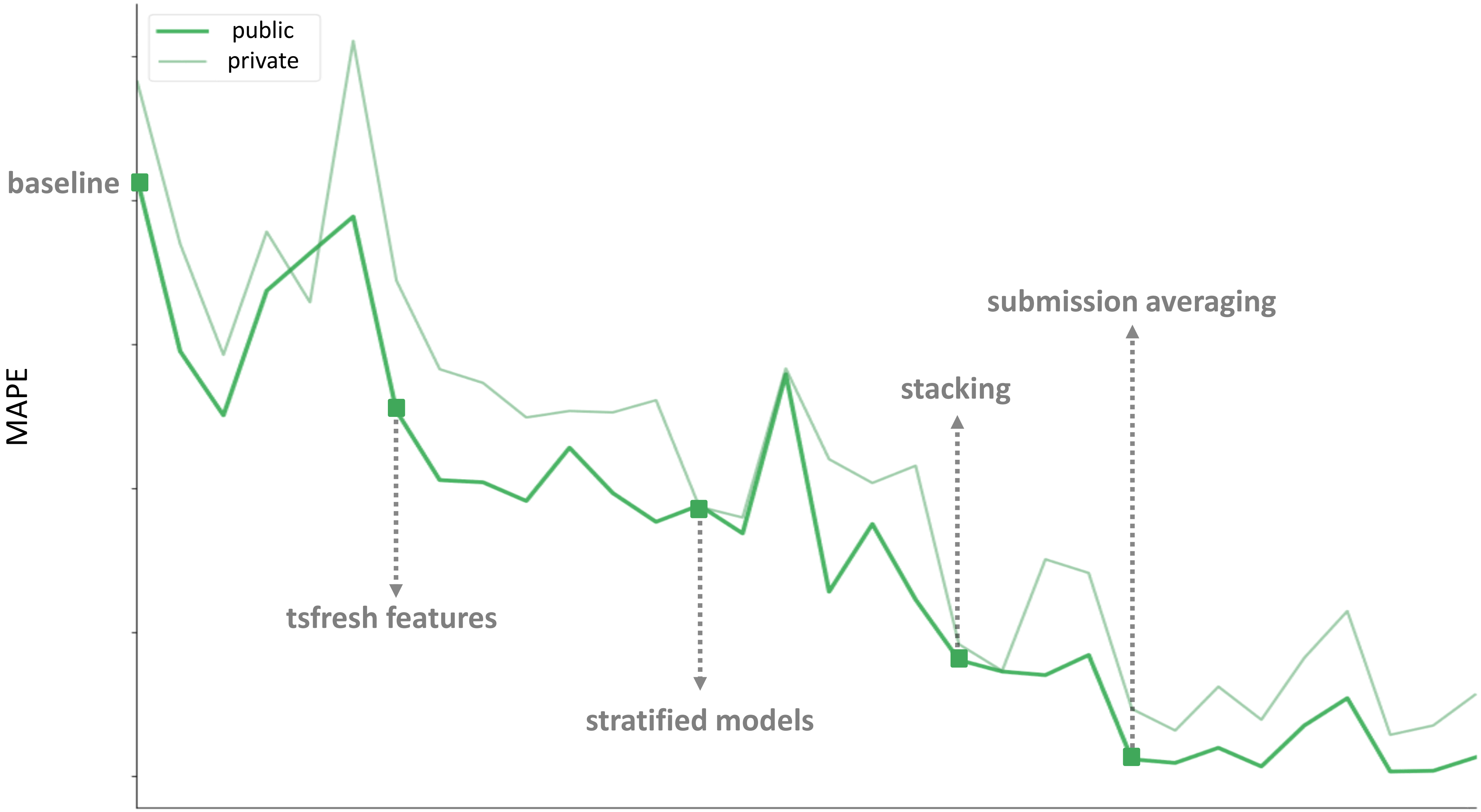
Team Fatima Yamaha

Private Score: 0.2658 (2nd Place)

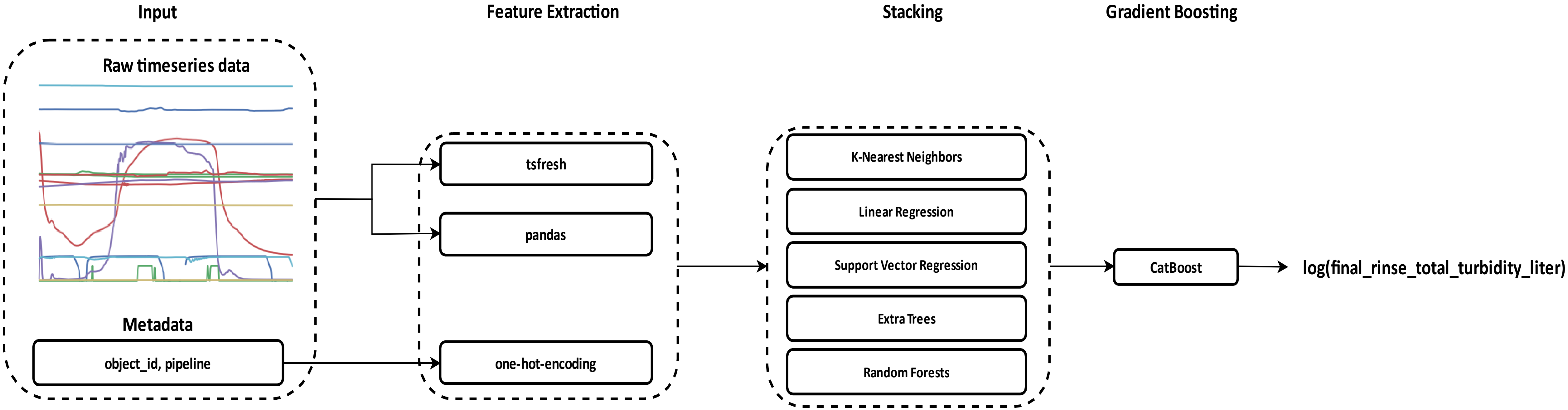
Gilles Vandewiele
PhD student

Thomas Mortier
PhD student

Starting with a simple baseline, using useful features and smart stratification, and ending up with a complex high performant ensemble, secured us the second spot



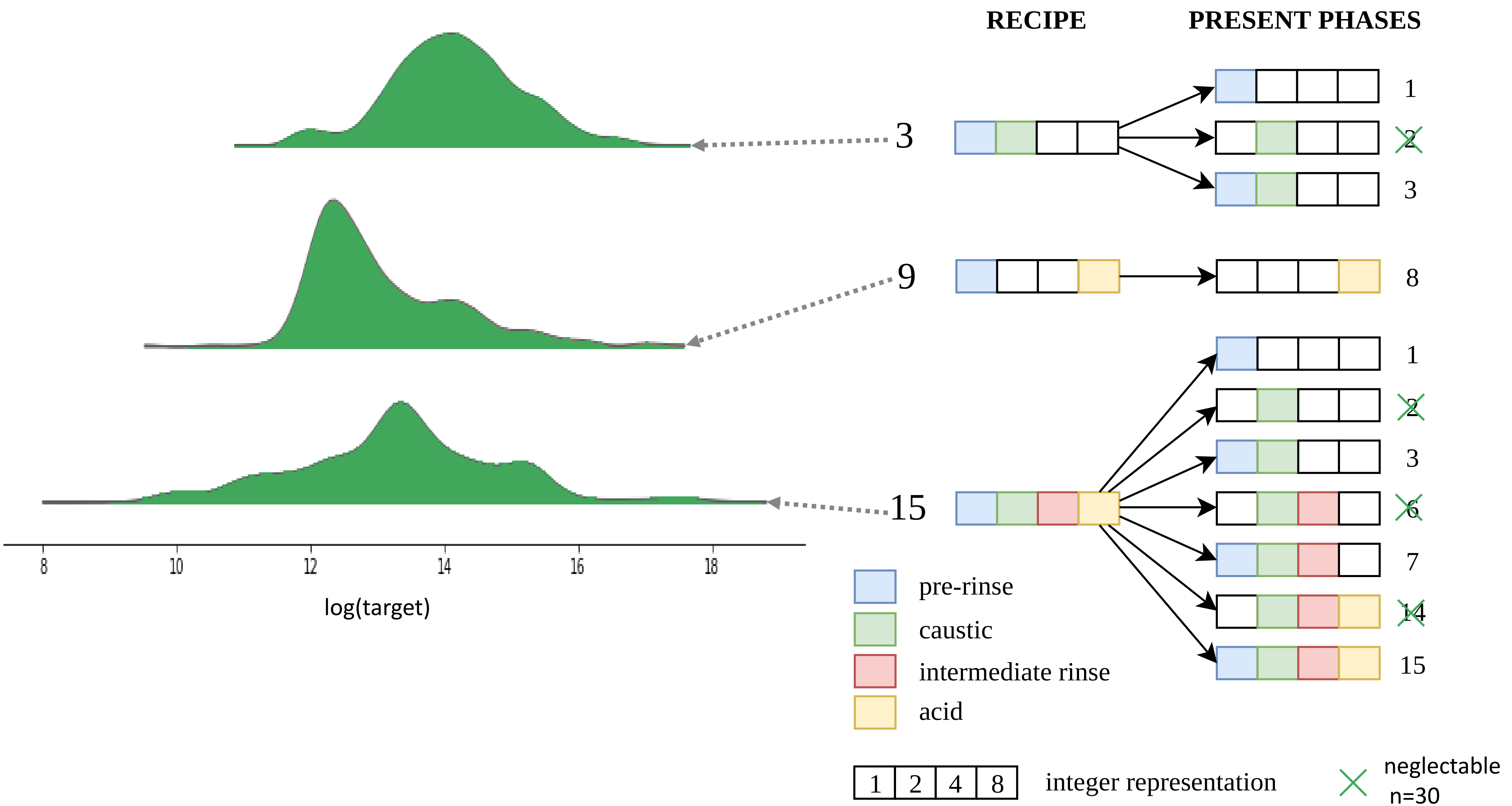
An overview of our final pipeline



CatBoost (https://github.com/catboost/catboost)

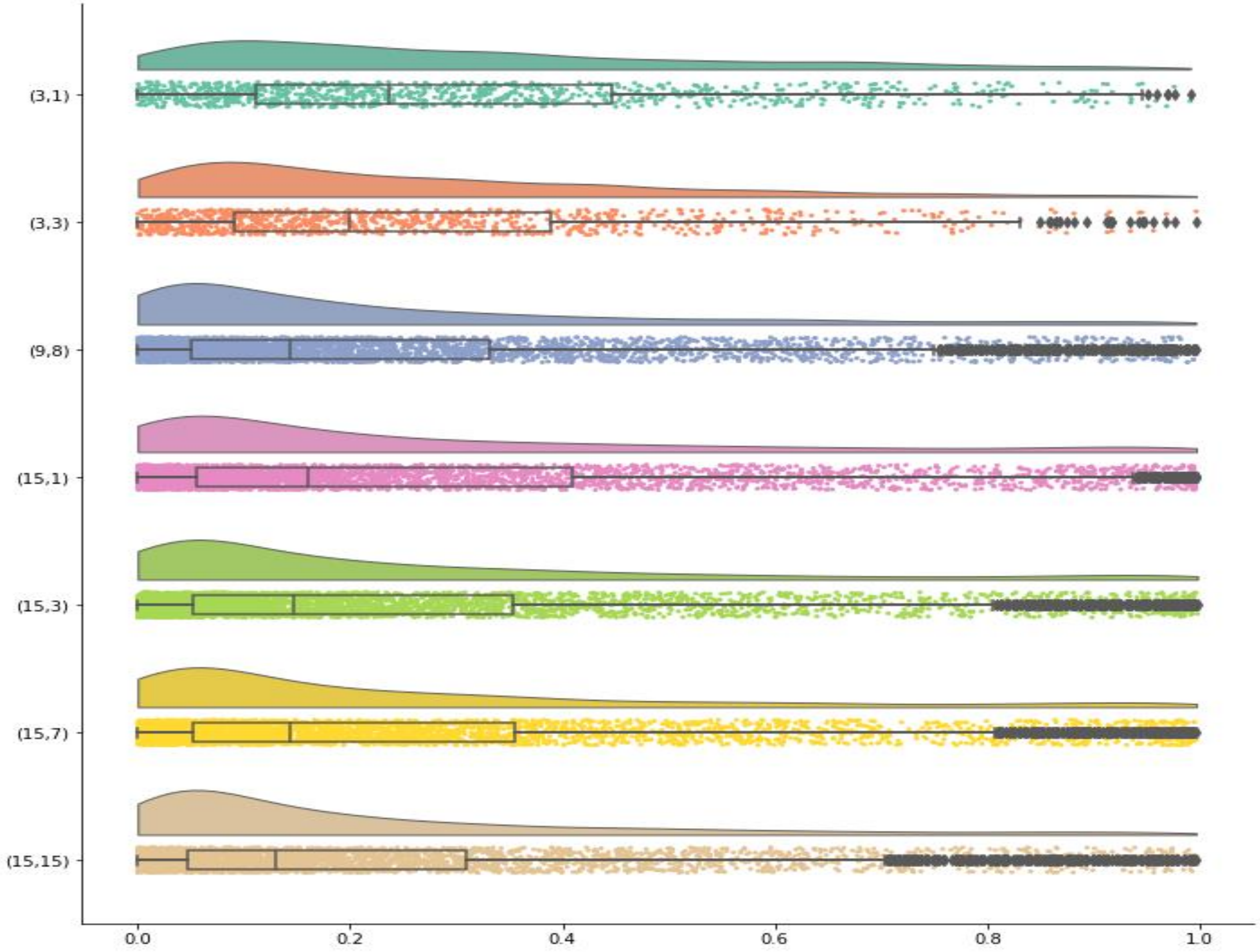
tsfresh (https://github.com/blue-yonder/tsfresh)

Stratifying models allow to predict after every cleaning phase and explains differences in variance in the outcome of the three cleaning recipes



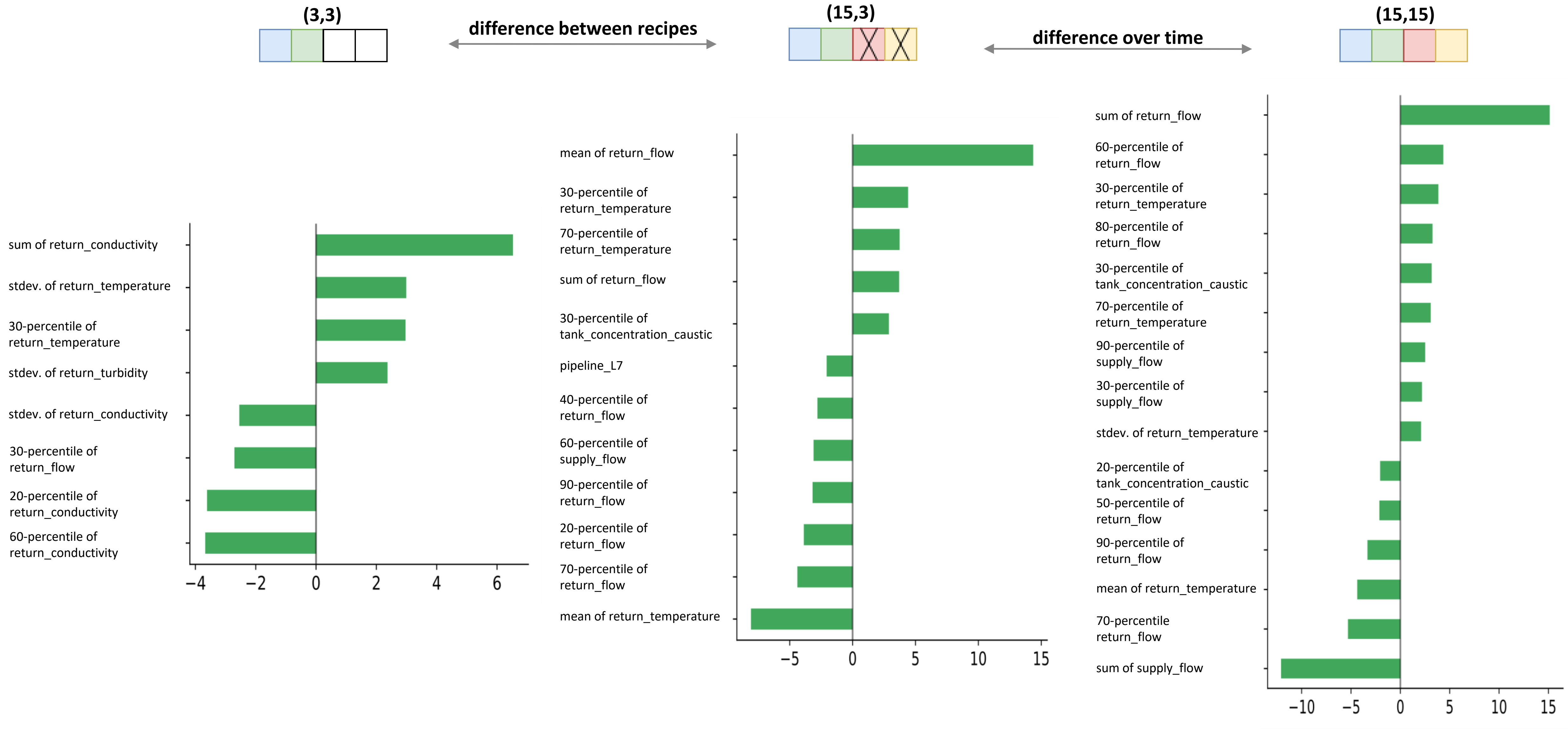
A higher number of processes and using more phases per processes decreases the five-fold cross-validation error, and reveals very high variance, hence, the choice for ensemble techniques

(*) the fraction of test processes **(**) model without stacking**



Group	MAPE $(\mu \pm \sigma)$	Weight*
(3,1)	0.35 ± 0.36	0.022
(3,3)	0.29 ± 0.32	0.17
(9,8)	0.38 ± 0.33	0.04
(15,1)	0.30 ± 0.37	0.07
(15,3)	0.28 ± 0.36	0.23
(15,7)	0.27 ± 0.34	0.22
(15,15)	0.25 ± 0.37	0.22
Total: **	$\mathbf{0.29 \pm 0.36}$	

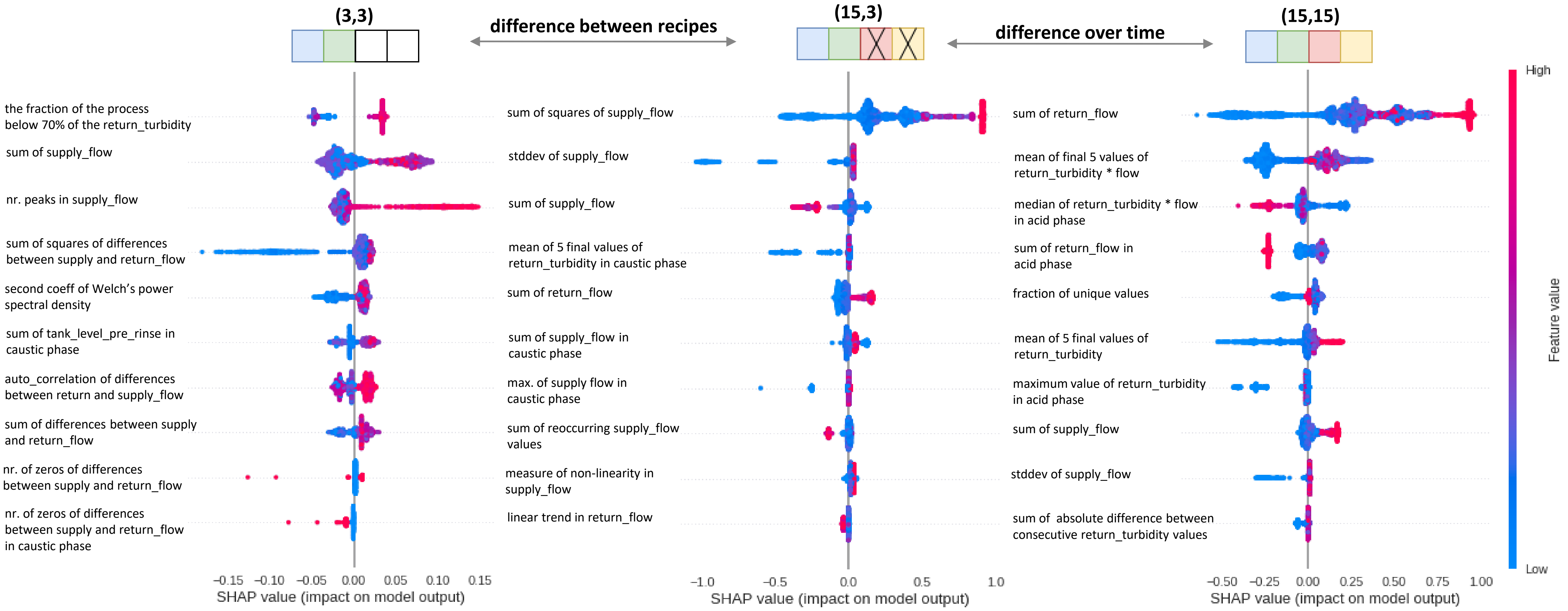
By using logistic regression, together with a p-value correction by means of the Benjamini-Hochberg correction (FDR, $\alpha = 0.05$), we identified positive and negative significant features which explain high targets



Shapley values allow to highlight most impactful features for the final rinse outcome

INTERPRETATION:

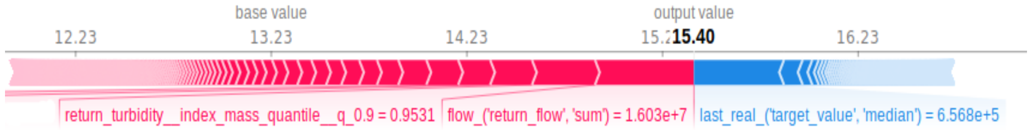
- dots along the x-axis are the different impacts on the model (left = large negative impact; right = large positive impact)
- color of dot impact feature magnitude (blue = lowest value, red = highest value)
→ e.g. for (3, 3), a higher number of peaks in supply_flow increases the prediction



shap (https://github.com/slundberg/shap)

Shapley values can be generated for each individual prediction and can be grouped together to find clusters of similar predictions and model behavior

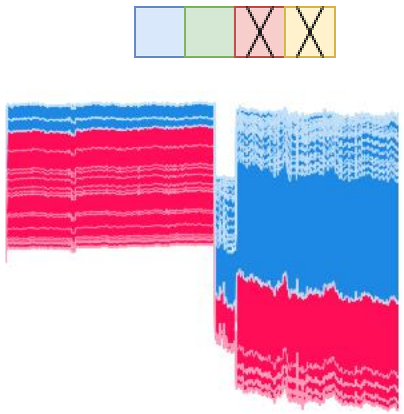
individual Shapley plot



rotate & cluster

(15,15)

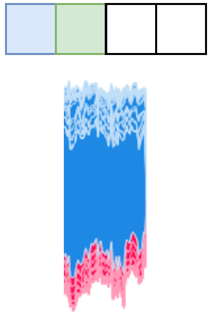
(15,3)



~ 300 predictions around 2.25 million liters due to high non-linearity in supply_flow

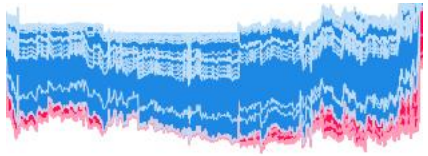
~ 300 predictions around 1.25 million liters due to low non-linearity and high variance in supply_flow

(3,3)

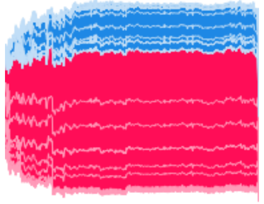


~ 50 predictions around 750,000 liters due to objects 205-209

~ 600 predictions around 800,000 liters due to relatively large sum of return_flow

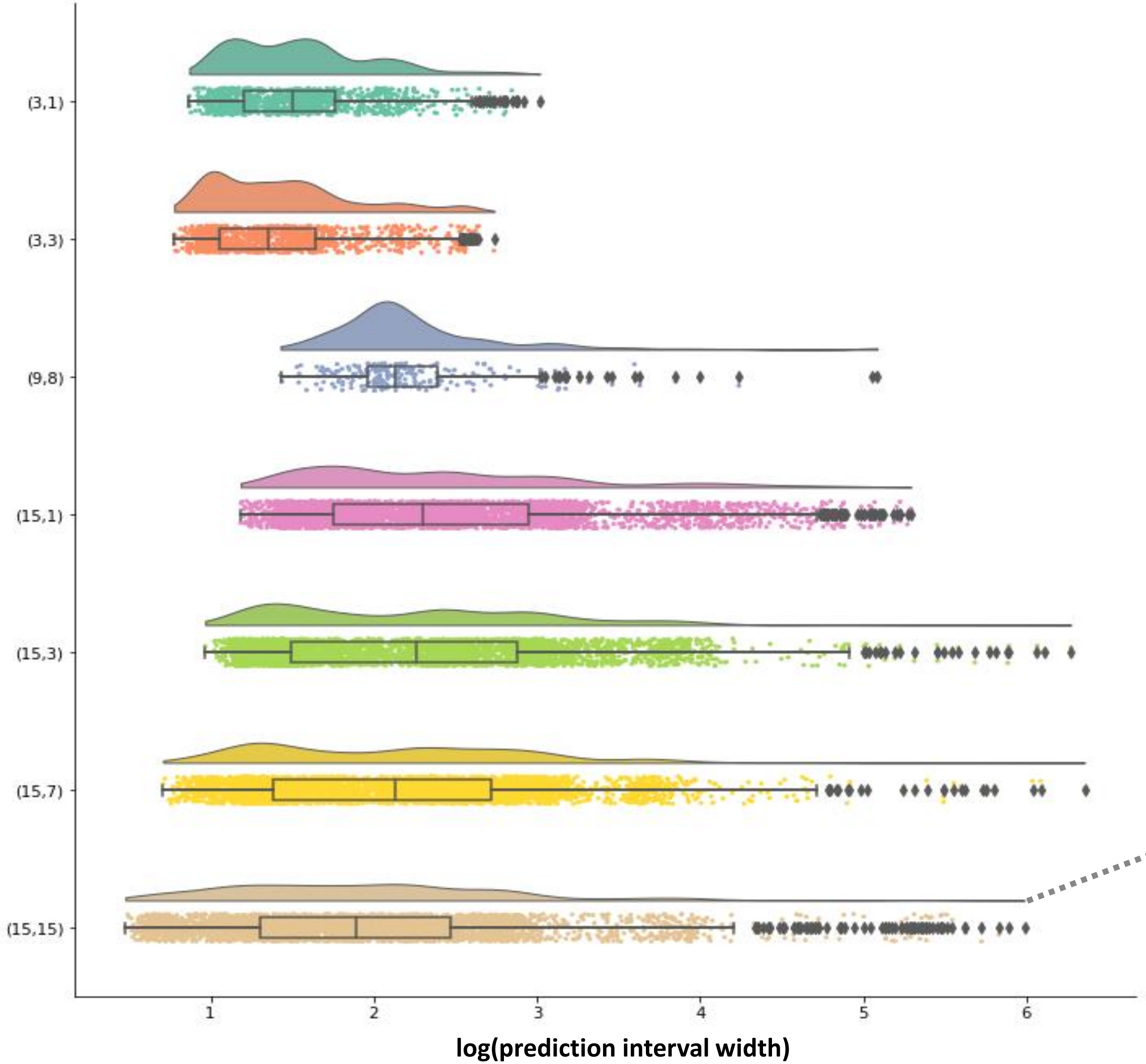


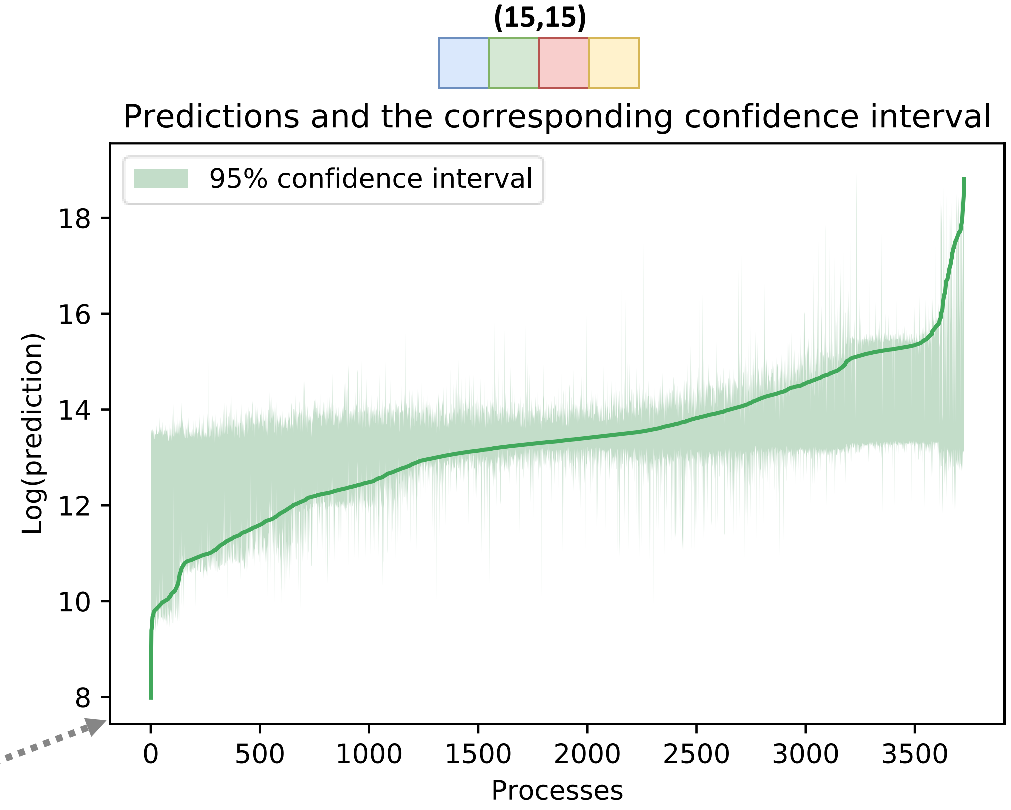
~ 400 predictions around 150,000 liters due to low sum of return_flow and high variance in supply_flow



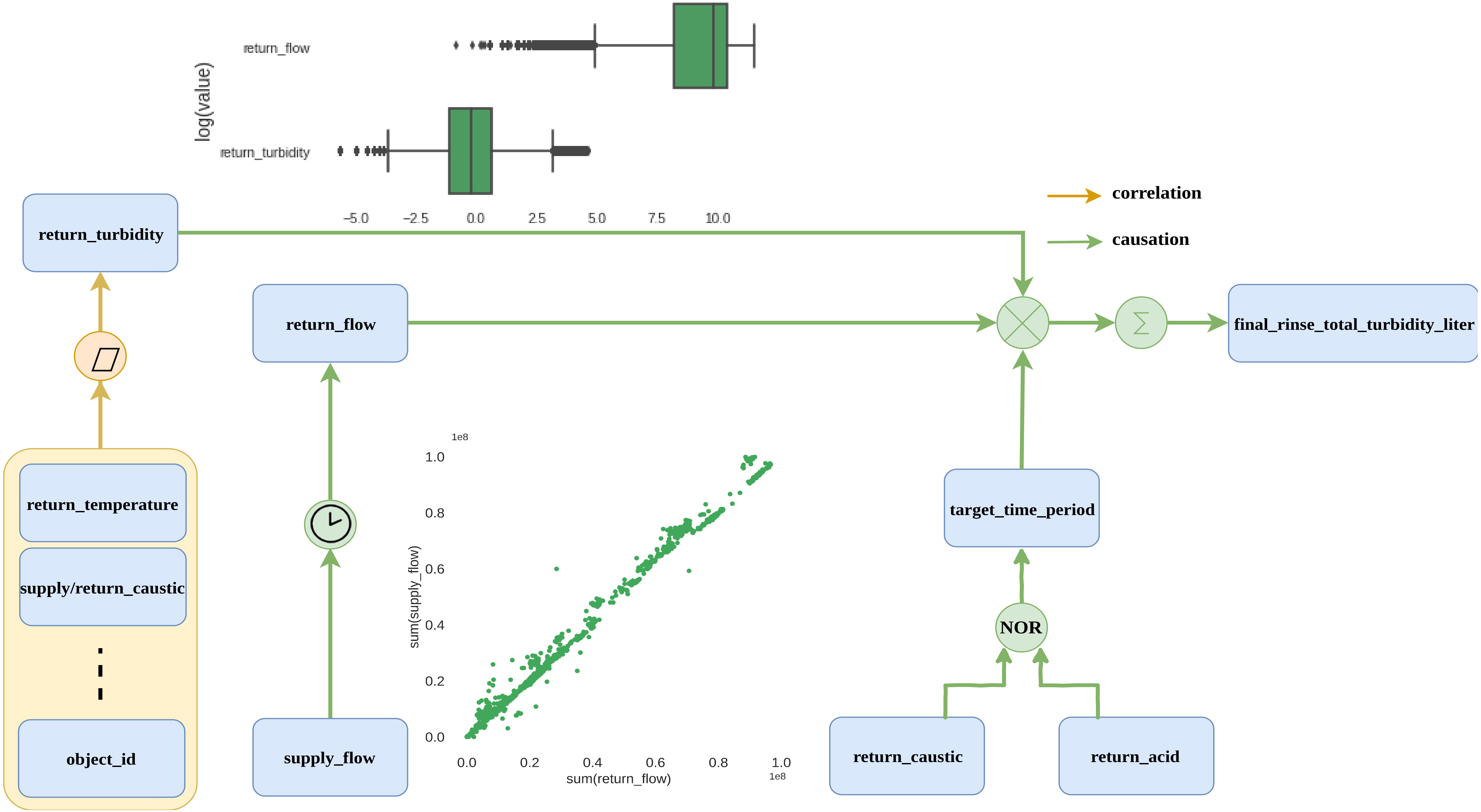
~ 350 predictions around 2 million liters due to much higher return_flow and supply_flow sums

Training two gradient boosters with quantile loss (1 - $\frac{\alpha}{2}$ and $\frac{\alpha}{2}$) allows to construct (1 - α)-prediction intervals with $\alpha = 0.05$





An aggregated target causes the less controllable return turbidity to be masked by return flow, which is of a larger order of magnitude while less informative



Predicting the turbidity outcome of the final rinse seems a harder task than predicting the flow outcome, however, interesting interactions appear

return_turbidity

Group	MAPE $(\mu \pm \sigma)$
(3,3)	0.33 ± 0.40
(9,8)	0.54 ± 0.49
(15,15)	0.33 ± 0.46
Total:	$\mathbf{0.34 \pm 0.45}$

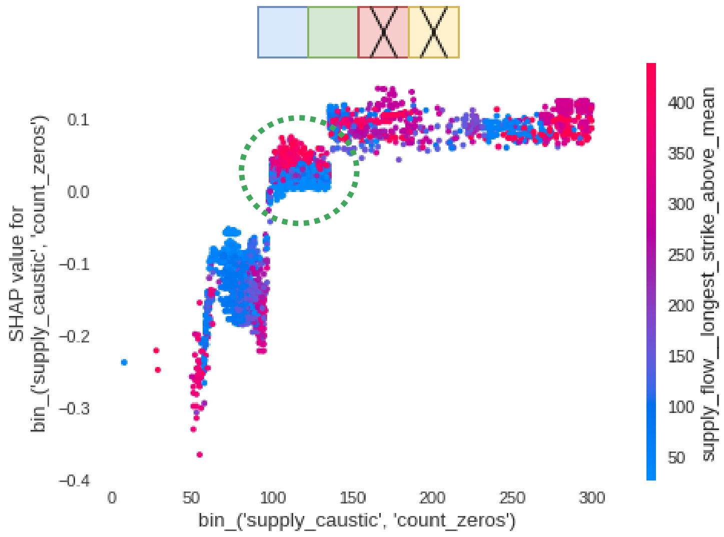
return_flow

Group	MAPE $(\mu \pm \sigma)$
(3,3)	0.11 ± 0.15
(9,8)	0.16 ± 0.17
(15,15)	0.10 ± 0.15
Total:	$\mathbf{0.10 \pm 0.15}$

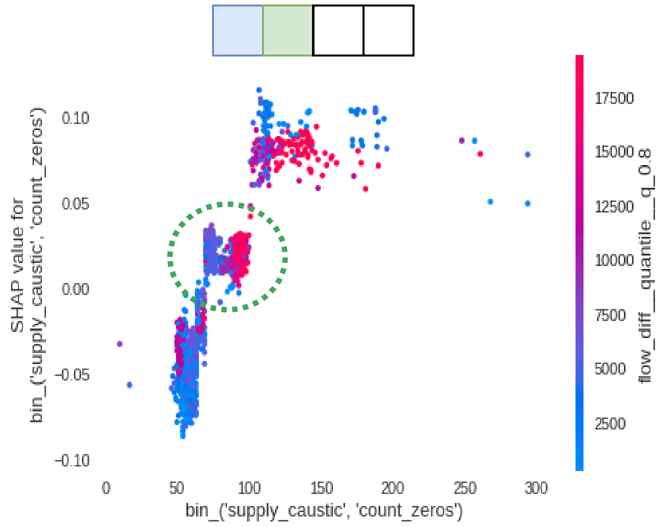
use only 22 flow features

Group	MAPE $(\mu \pm \sigma)$
(3,3)	0.10 ± 0.16
(9,8)	0.19 ± 0.21
(15,15)	0.11 ± 0.17
Total:	$\mathbf{0.11 \pm 0.17}$

(15,3)

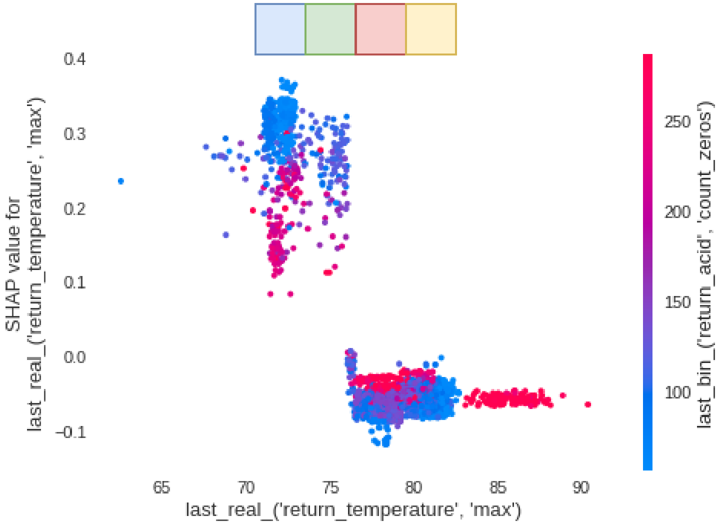


(3,3)



three clusters can be identified, for both recipes, after the caustic phase, based on nr. zeros in supply_caustic, whereas within the highlighted clusters another partition can be made

(15,15)



when the maximum return_temperature exceeds 76 °C the return_turbidity from the final rinse will be lower, moreover within low-temperature processes, a new partition can be made depending on the inactivity duration of the acid valve