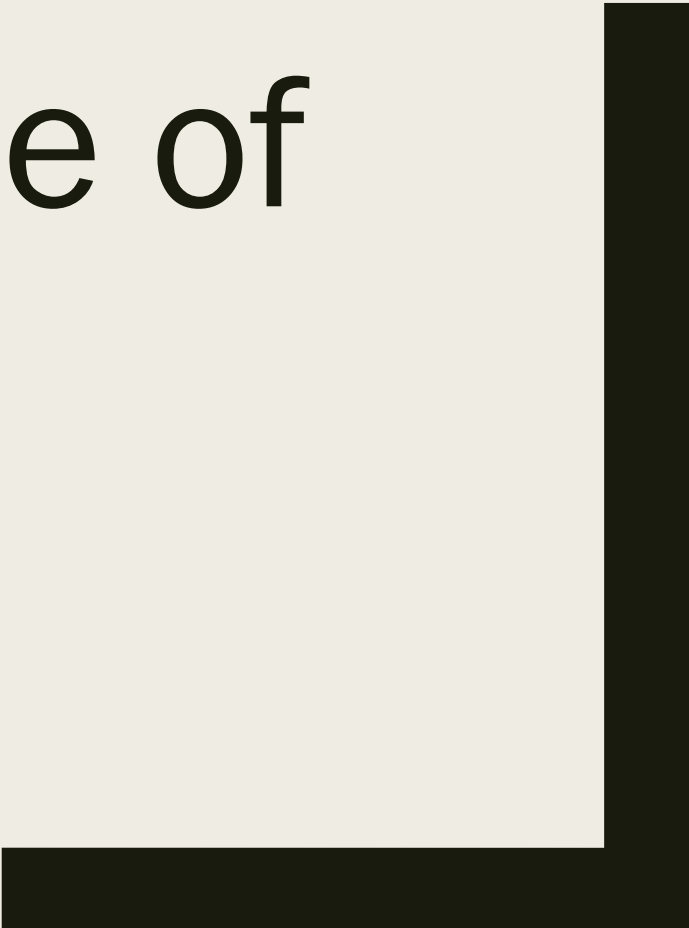




The importance of the object

mlearn
March 2019

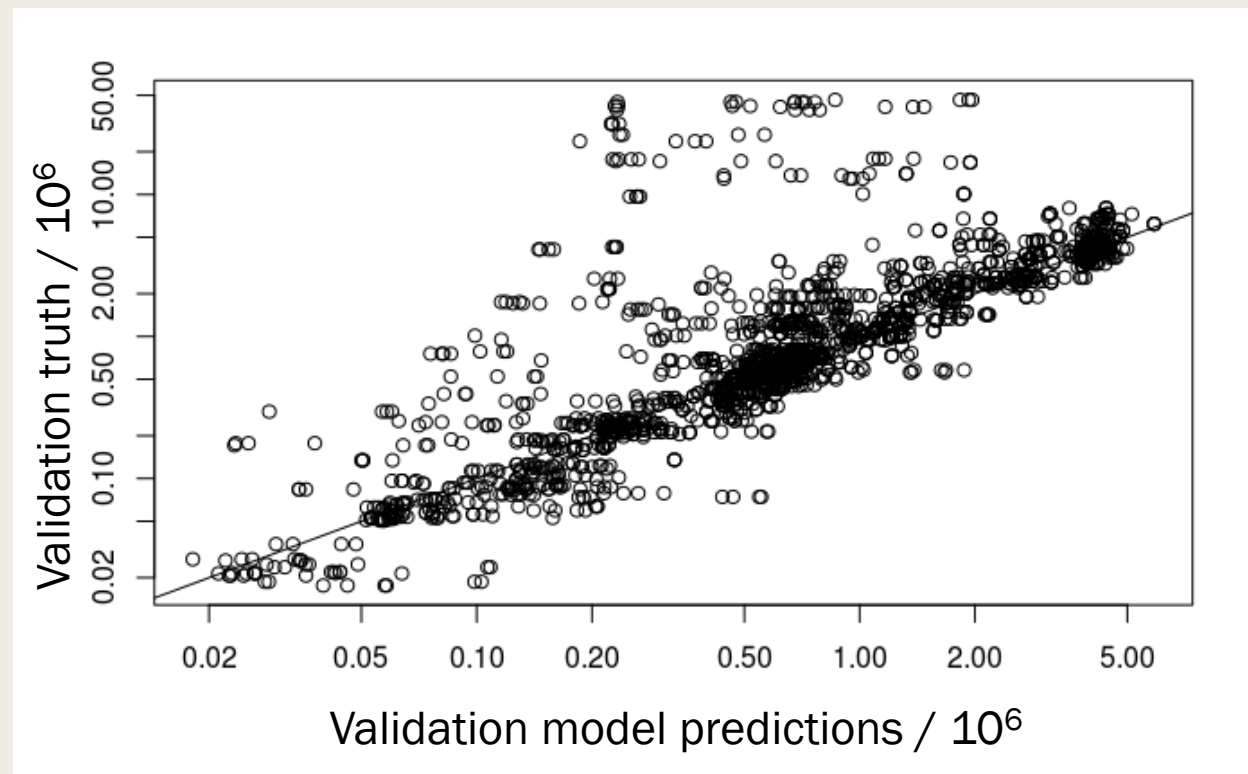


Introduction

- I modelled final turbidity as a function of clean-in-place measurements
- Complex machine learning is not adding much as a predictive model
 - *A simple model based on average object history gets a mean absolute percentage error of 30.4%*
 - *My competition model improves this slightly to a mean absolute percentage error of 26.5%*
 - *The competition winners may be a couple of percentage points better off again*
 - *These competition models do not give a predictive model performance gain worth large implementation complexity*
- I analyse our model and related models. Various patterns can be seen but they are not useful patterns for advancing the clean-in-place industry
 - *The object ID is dominant and we have no object characteristics to add any physical understanding*
 - *I suspect that most simple patterns in the data are likely to be a proxy for the object perhaps due to human or computer process control*

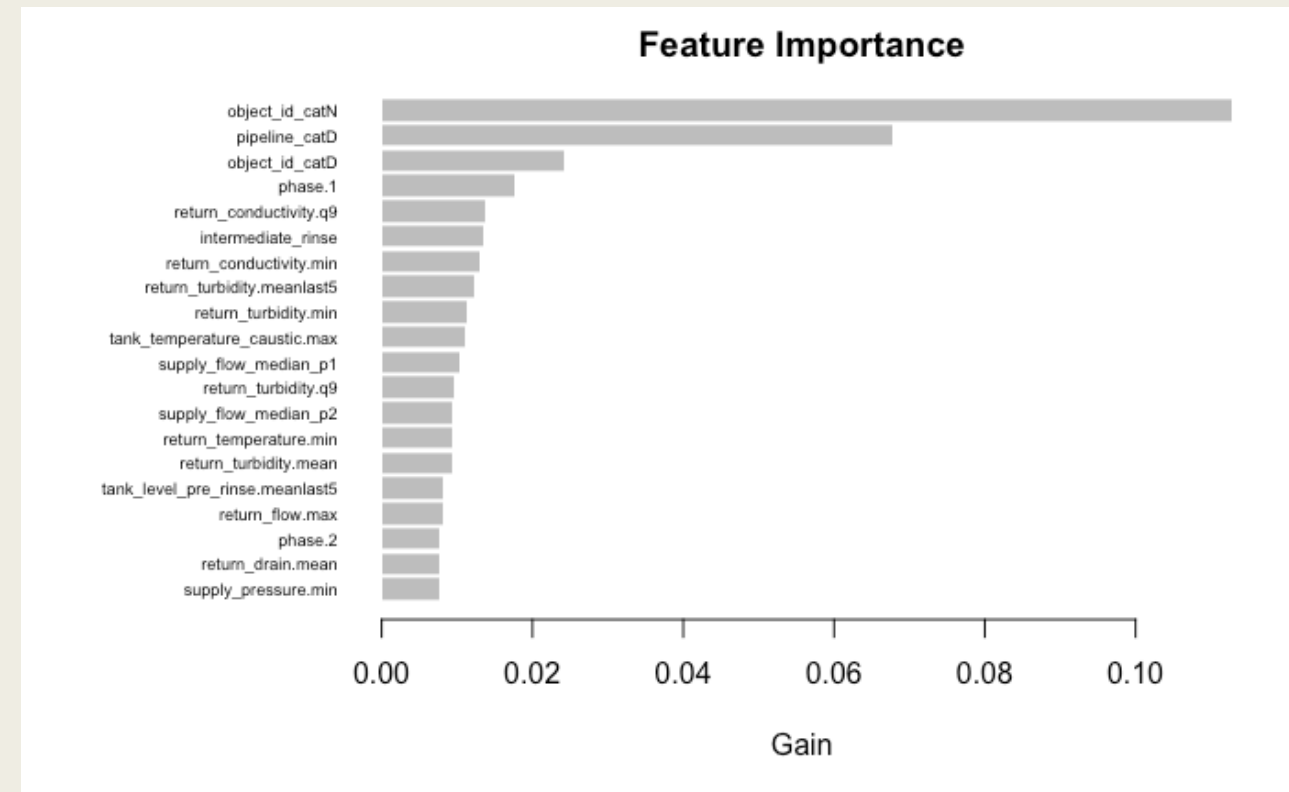
Summary of my modelling approach

- Extract 249 features mainly from individual time series
 - Features will be listed as “*timeseries_summary*”, e.g. *supply_flow_q1* for 1st decile of supply flow
- Data augmentation to match the test data generation process
- Fit gradient boosted decision trees (lightgbm)
 - Can include the competition loss function
- Details in submitted model
- Used 90:10 training:validation split
 - Validation scores are quoted in this presentation
 - On the right the fit is shown for the validation set



Variable sensitivity

- Calculate decision tree Gini gain per feature (shown on the right)
- The gain is dominated by the object ID (top three features):
 - *catN and catD are categorical item codings by the R package vtreat*
 - *catN is the shift in the mean response for a categorical level*
 - *Objects only appear on a single pipeline so pipeline_catD represents a group of objects*
- So much is potentially wrapped up in the object ID, e.g. type and size of object and typical dirtiness. We can go no further with this competition data.
- **Recommendation 1:** Provide more object information. This may allow understanding to be derived about real-world properties of objects. Further this will allow any predictive model to handle new objects and probably better handle rare objects

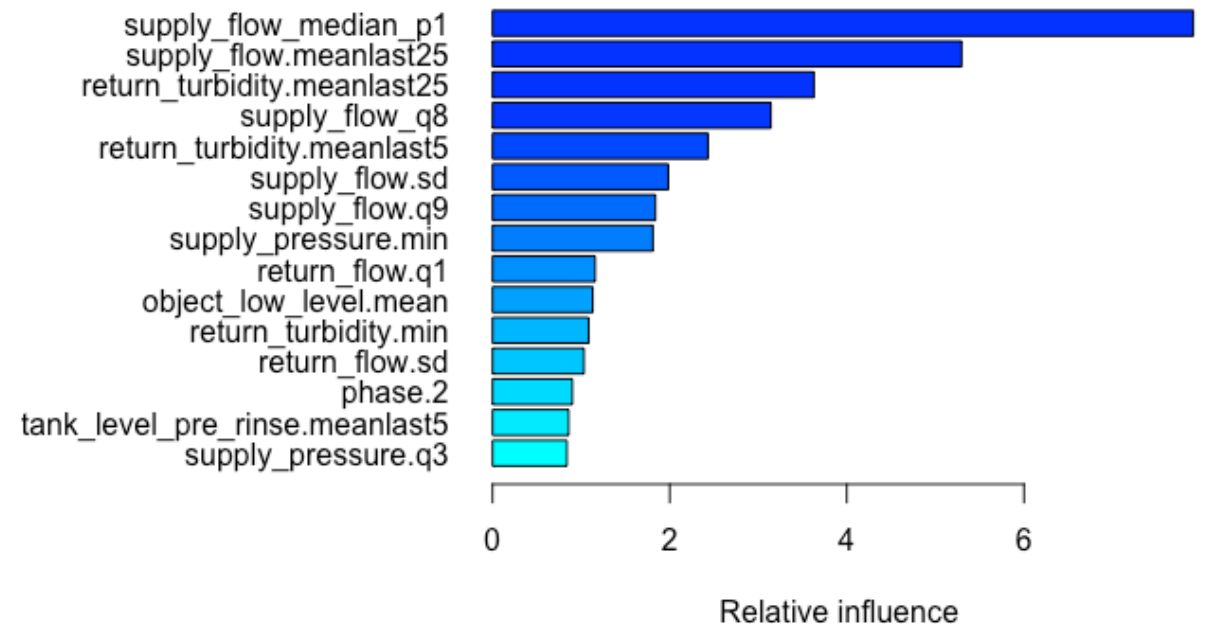


Object-only model

- We have seen how dominant the object ID is. Let us now see how powerful a model based on just object ID is.
- I built a simple small gbm model just using the object ID and quickly achieved a validation error of 0.304 (i.e. a mean absolute percentage error, MAPE, of 30.4%)
 - *I imagine this error could be improved on with more effort, either via parameter tuning or perhaps changing to a Bayesian model*
- It is not clear that a more complex model such as my best model (that achieves a validation error of 0.265, 26.5% MAPE) is particularly more useful
- **Recommendation 2:** If a predictive model is needed in the field then focus on average object history which should give a simple and easy-to-maintain model

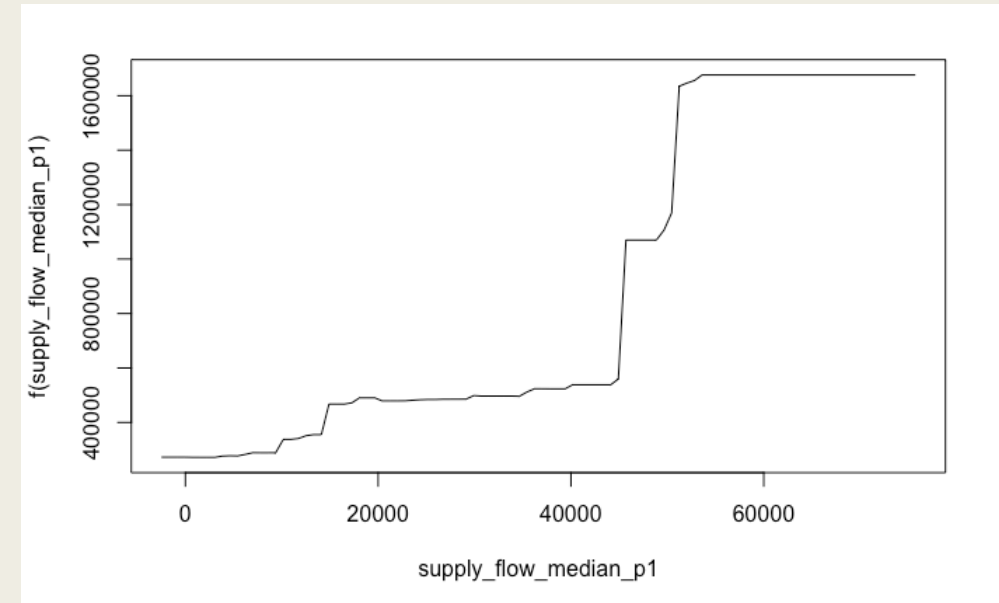
No object ID: Model

- We will try to understand process features by removing the object features from our modelling
 - *The resultant model is not great but is competitive (MAPE 30.1%)*
- The Gini feature importance plot (right) initially looks better
 - *supply_flow and return_turbidity seem plausible important measures*



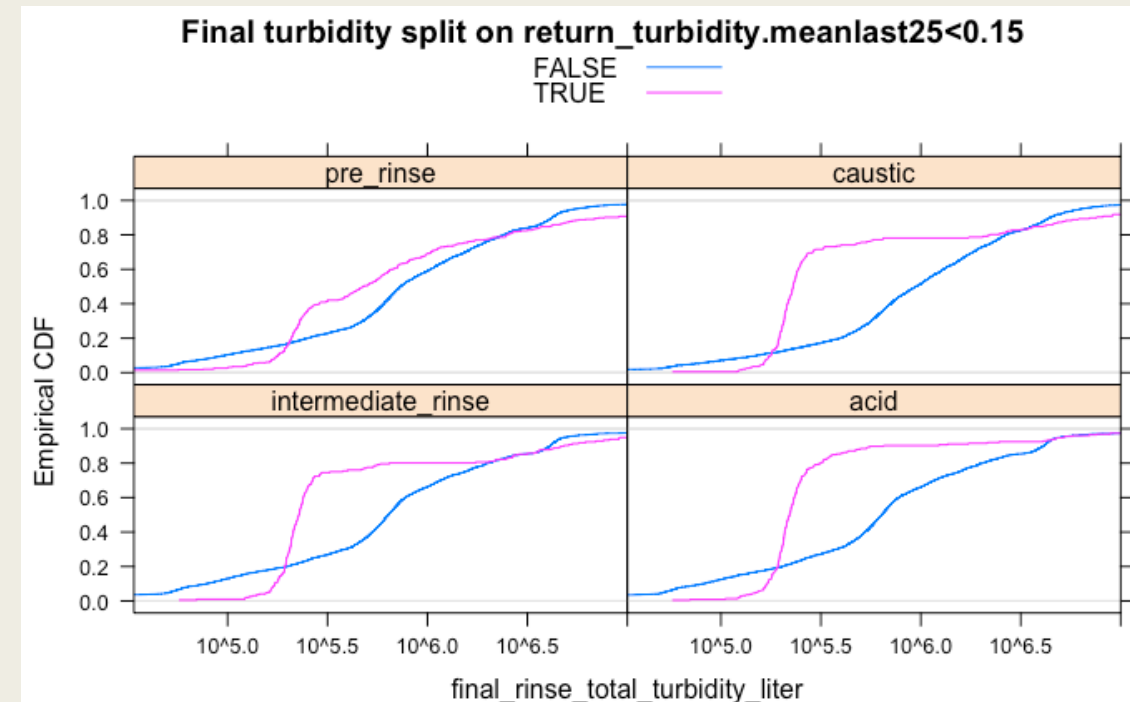
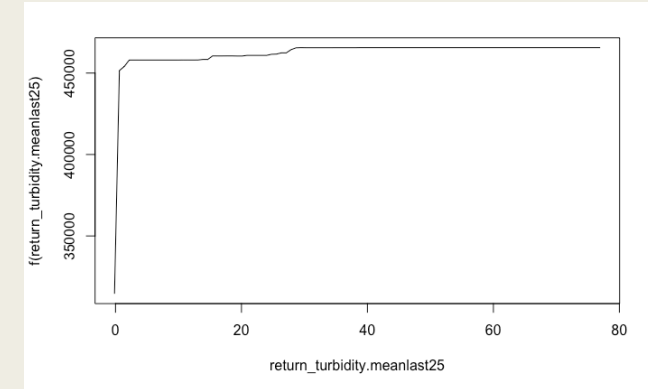
No object ID: Supply flow is not a useful predictor

- The dominant feature in the no-object-ID model is the median of the supply flow in the pre-rinse stage. The partial dependency plot shows a large roughly monotonic increase (right).
 - *Other supply flow features are also generally monotonically increasing but weaker*
- This monotonicity is the opposite direction than one might naively believe – one might expect more flow to lead to a cleaner outcome
- Is supply flow an independent variable set by a computer or a human based on seeing the object? Is more flow used for, say, larger objects which require more cleaning and which still end up with more residual turbidity?
 - *This hypothesis is supported by a strong correlation between supply flow and object ID (e.g. Pearson's r^2 between `supply_flow_median_p1` and `object_id_catN` is 0.54)*
- We are left assuming the supply flow is being used by the model as a proxy for object ID.
- **Recommendation 3:** Consider a statistical experimental design or randomised experiment to collect more diverse data. This will allow the model to understand the effects of different process settings (in this case supply flow).



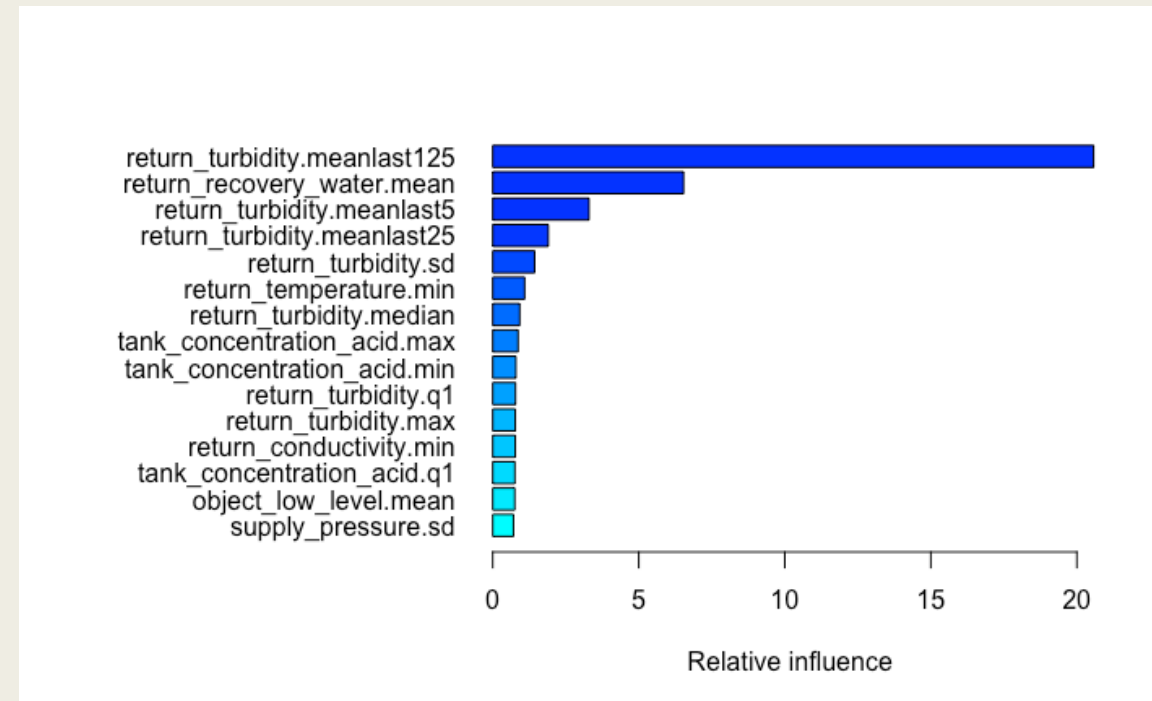
No object ID: Return turbidity is a weak signal of final cleanness

- In the no-object-ID model we also see the return turbidity over the last 25 measurements as important
- Recent turbidity measurements are unsurprisingly slightly predictive of final turbidity measures
- The strongest feature is the mean over the last 25 measurements
 - *If we've already seen clean flow then likely to remain clean otherwise no signal (partial dependence plot top-right)*
- This does suggest a weak criteria for predicting final cleanness
 - *The empirical cumulative distribution functions of final turbidity based on a 0.15 decision threshold on the mean of the last 25 observations of return turbidity in each phase is shown on the bottom-right. After the pre-rinse stage the difference is noticeable and this decision threshold is particularly reliable at the end of the acid phase*
- **Recommendation 4:** Consider shorter washes on cases when return_turbidity is low



Single object model: another way of convincing us of lack of signal

- I also looked at building models per object to try to remove the strong object ID effect (direct or indirect from other features)
- Did not find a particularly strong relationship to further features
- Examples for three common objects:
 - *Models on objects 405 or 933 failed to find strongly predictive features*
 - *Object 932 gets a model that has some predictive power but is dominated by recent return_turbidity measurements (right) – this does not tell us anything new*



Conclusions

- Machine learning with many features can eke some signal out of this data beyond object history. This signal is weak and is not a result of a small number of features.
- If we knew more about the object that may allow association with physical properties (recommendation 1)
- Object ID can be used to form a strong predictive model without any other features (recommendation 2)
- A controlled experimental design may be required to advance the field (recommendation 3)
- Return turbidity should be monitored and if very low this can potentially be used as a signal for a shorter wash (recommendation 4)