Article:

Mounce, S.R. orcid.org/0000-0003-0742-0908, Ellis, K., Edwards, J.M. et al. (3 more authors) (2017) Ensemble decision tree models using RUSBoost for estimating risk of iron failure in drinking water distribution systems. Water Resources Management, 31 (5). pp. 1575-1589. ISSN 0920-4741

https://doi.org/10.1007/s11269-017-1595-8

Ensemble decision tree models using RUSBoost for estimating risk of iron failure in drinking water distribution systems

Supplementary Material 2 – additional data analysis

The table below provides a % missing data figure for all variables. In addition, the correlation r value between an individual variable and the iron failure target is given. Of course, this is only a bivariate value but gives a sense of the potential of candidate predictor variables.

Sampling frequency and failure were not used (other than iron failure for the target) since these are activity/ reactive based and to avoid a circular self-fulfilling system (the more sampling carried out, the more likely a failure is to be found – also because sampling across parameters is generally done at the same time so these are dependent). Any variable with >60% missing data was removed (all of these over 90% missing data except for SRV_IRON_AV which is a promising variable but has 63.7% missing data) – these are indicated in red. Some of the other SRV and WTW parameters could have a beneficial effect on model accuracy if additional data can be collected in the future.

TreeBagger bags an ensemble of decision trees for either classification or regression. Bagging stands for bootstrap aggregation. Every tree in the ensemble is grown on an independently drawn bootstrap replica of input data. Observations not included in this replica are "out of bag" for this tree. Treebagger can be used to assess feature importance for classification (see Loh, W.Y. and Y.S. Shih. "Split Selection Methods for Classification Trees." *Statistica Sinica*, Vol. 7, 1997, pp. 815–840.). Results for the top ten variables (in bold) are provided in the figure. These were ultimately used, except for WTW Chlorine total which proved the weakest variable on the Treebagger results and for which expert evaluation did not expect a correlation.

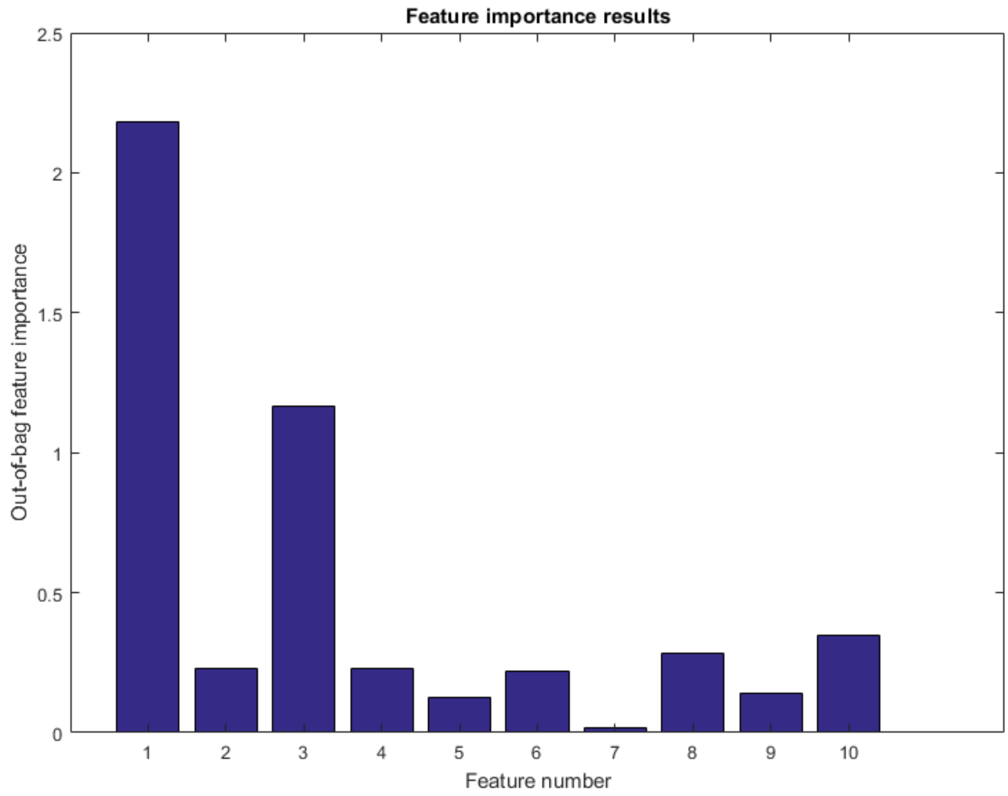
The three most important features from both methodologies are the average of median Iron measurements in a DMA, the average of median turbidity measurements in a DMA and the total number of customer contacts (complaints) about water quality.

Data Matrix Statistics

Parameter name	**% missing data**	**Pearson Correlation with Iron fail target (H=1, L=0)**
dma	0%	N/A
wqz	0%	N/A
population	0%	-0.011
iron_Nsamples	0%	N/A
iron_Nfails	36.5%	Converted to target output
iron_av	36.5%	**0.372**
mang_Nsamples	0%	N/A
mang_Nfails	44.0%	N/A
mang_av	44.0%	**0.074**
turb_Nsamples	0%	N/A
turb_Nfails	47.8%	N/A
turb_av	47.8%	**0.318**
chlor_total_Nsamples	0%	N/A
chlor_total_av	38.0%	-0.082
chlor_free_Nsamples	0%	N/A
chlor_free_av	38.7%	-0.070
pH_Nsamples	0%	N/A
pH_av	42.5%	-0.070
temp_Nsamples	0%	N/A
temp_av	38.0%	-0.035
cc	42.9%	**0.187**
cc_clusters	DERIVED NOT IN ORIGINAL DATASET	**0.128**
iron_lined	7.2%	**0.031**
iron_unlined	7.2%	**0.059**
other_material	7.2%	0.021

Parameter name	**% missing data 2008-2014**	**Pearson Correlation with Iron fail target (H=1, L=0)**
total_length	7.2%	**0.042**
wtw_iron_Nsamples	0%	N/A
wtw_iron_av	91.5%	-0.058
wtw_mang_Nsamples	0%	N/A
wtw_mang_av	90.9%	-0.079
wtw_turb_Nsamples	0.5%	N/A
wtw_turb_av	4%	**0.011**
wtw_chlor_total_Nsamples	0.5%	N/A
wtw_chlor_total_av	4%	**0.057**
wtw_chlor_free_Nsamples	0.5%	N/A
wtw_chlor_free_av	4%	0.011
wtw_pH_Nsamples	0%	N/A
wtw_pH_av	91.6%	0.013
wtw_temp_Nsamples	0.5%	N/A
wtw_temp_av	3.9%	-0.022
srv_iron_Nsamples	0.03%	N/A
srv_iron_av	63.7%	0.131
srv_mang_Nsamples	0.03%	N/A

Parameter name	% missing data 2008-2014	Pearson Correlation with Iron fail target (H=1, L=0)
srv_mang_av	94.0%	0.153
srv_turb_Nsamples	0.03%	N/A
srv_turb_av	99.3%	0.259
srv_chlor_total_Nsamples	0.03%	N/A
srv_chlor_total_av	16.3%	0.038
srv_chlor_free_Nsamples	0.03%	N/A
srv_chlor_free_av	16.3%	0.020
srv_pH_Nsamples	0.03%	N/A
srv_pH_av	98.5%	-0.009
srv_temp_Nsamples	0.03%	N/A
srv_temp_av	16.3%	-0.033
Nflushes_routine	57.1%	0.046
Nflushes_reactive	99.9%	0.079
Nbursts	28.6%	0.018



1. iron_av
2. mang_av
3. turb_av
4. iron_unlined
5. iron_lined
6. total_length
7. wtw_chlor_total_av
8. wtw_turb_av
9. cc_clusters
10. cc