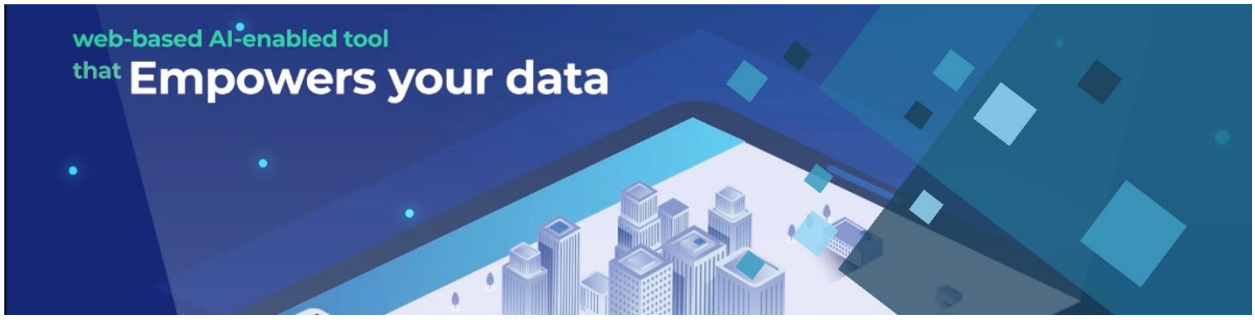

infraPLAN Ask the Experts! #3

Break Prediction of infraSOFT Machine Learning Module up to 6 times Superior to Desktop Scoring

Case Studies

In a previous article (#2) we have shown that assessing the Likelihood of Failure (LOF) of a pipe using *desktop* or *GIS scoring* can be difficult to complete and limited in scope; and that relying on advanced analytics is easier to develop.

In this article we compare the break prediction capacity of *desktop scoring, multi-variable regression,* and *machine learning* when applied to three systems of different size, physical condition, data quality, and break history.

We show that the capacity of our proprietary *machine learning* model - embedded in our infraSOFT platform- to predict breaks is up to 6 times superior to *desktop scoring*; up to 90% of the breaks occurring during the next two years can be predicted. Our *machine learning* model performs extremely well even on a small system, or a system with an erratic break trend, or problematic data.

info@infraplan-llc.com www.infraplan-llc.com

Project #1 - Large system - Medium break rate

1,490.0 miles of Cast Iron (CI) pipes and 713.6 miles of Ductile Iron (DI) pipes

The validation technique is described in this first case study; it is the same regardless of the modeling approach.

Project #1 took place in 2022 with break data from 2002-2021. Breaks from 2002-2019 are used to generate a predicted LOF with each modeling approach considered -for this project, *desktop scoring* and *machine learning*- for 2020-2021 for each pipe.

The pipes are then ranked based on their LOF, with the highest LOF first. For each pipe we also have the actual breaks that occurred in 2020-2021. If a pipe had been replaced by 2020, its 2020-2021 breaks would have been avoided. We therefore, consider a certain percentage of pipes (ranked with highest LOF first) and compute the percentage of the total number of breaks experienced in 2020-2021; they would have been avoided had those pipes been replaced by 2020.

The higher the percentage of breaks avoided during the validation period, the better the model.

The Table below shows the validation results obtained with *desktop scoring* and *machine learning* for Project #1. *Multi-variable regression* was not used for this project.

ALL PIPES	% Breaks (2020-2021) avoided (worst pipes first)		
% Pipes Targeted	**Desktop scoring**	**Multi-variable regression**	**Machine Learning**
1	17.8	NA	30.6
5	**19.4**	NA	**79.3**
10	**45.2**	NA	**84.9**

If the top **5%** worst pipes as ranked by their *desktop scoring* LOF had been replaced by 2020, **19.4%** of the 2020-2021 breaks would have been avoided; if ranked with m*achine learning* LOF scores, that percentage becomes **79.3%.** For the top **10%** worst pipes, the percentages are **45.2%** (desktop) versus **84.9%** (ML). The performance of the *machine learning* model is excellent with this data set.

For Project #1 the capacity of our machine learning model to predict breaks is 1.8 to 4 times better than desktop scoring.

info@infraplan-llc.com www.infraplan-llc.com

Project #2 - Very Large system - High Break Rate

2,457.7 miles of Cast Iron (CI) pipes and 2,511.7 miles of Ductile Iron (DI) pipes

Project #2 took place in 2021 with 2005-2020 break data; 2005-2018 breaks are used to predict breaks for 2019-2020 which are compared with the actual 2019-2020 breaks.

The Table below shows the validation results obtained with *desktop scoring*, *multi-variable regression*, and *machine learning*. For this project a *multi-variable regression* model had to be created for each material because of calibration constraints that characterizes that approach. Therefore, to compare apples to apples, the *desktop scoring* and *machine learning* models were validated separately for each material. Each cell contains two values, first the percentage of breaks avoided by DI pipes, and then by CI pipes.

DI/CI	% Breaks (2019-2020) avoided (worst pipes first)		
% Pipes Targeted	**Desktop scoring**	**Multi-variable regression**	**Machine Learning**
1	0.6/1.1	34.0/12.1	44.2/15.3
5	**12.2/5.9**	**54.5/35.7**	**63.5/38.1**
10	20.2/9.2	69.4/52.0	72.0/55.2

Results show that, for example, if the top **5%** worst pipes as ranked with LOF scores generated with *desktop scoring*, *multi-variable regression*, or *machine learning*, had been replaced,
- **12.2%, 54.5%,** and **63.5%** of the 2019-2020 breaks (for DI pipes), and;
- **5.9%, 35.7%,** and **38.1%** of the 2019-2020 breaks (for CI pipes), respectively, would have been avoided.

For this project, *advanced analytics* also yield much better results than *desktop scoring*. However, the scale of the improvement depends on the percentage of pipes replaced, the material and approach.

CI pipes constitute a cohort for which predictions are more difficult to make regardless of the approach (the data needs further improvement). For those pipes, *desktop scoring* yields results that are slightly better than rolling a dice at 5%! However, even for that rather weak cohort, results are still around 6 times better at 5% and 10% replacement with *machine learning* than with *desktop scoring*.

Similarly to Project #1, Project #2 also illustrates that advanced analytics has a much higher failure forecasting capacity than desktop scoring, with machine learning performing slightly better than multi-variable regressio*n*. The good performance of the regression model is partly thanks to adequate calibration choices which have to be made manually by the user for every run. This can be tedious and requires expertise. While a strong *machine learning* model also requires that the specificity of pipe and break data be taken into account, this is embedded in the software; the model is internally programmed by the data scientist for automated calibration making it a more practical option for the user.

For Project #2 the machine learning break prediction model performs up to 6 times better than desktop scoring even for a data set with rather poor data; machine learning performs slightly better than multi-variable regression. It is also an approach superior to multi-variable regression because the burden of calibration does not rest on the user at each run.

.

Project #3 - Small system - Erratic break trend

15.5 miles of Cast Iron pipes and 187.7 miles of Ductile Iron pipes.

This study took place in 2020 with 2007-2019 break data. 2005-2017 breaks are used to predict breaks for 2018-2019 which are compared with the actual 2018-2019 breaks. Breaks have been erratic with an overall upward trend from 2007 to 2019. However, spikes were observed in 2010, 2012 and 2015 with up to 4 times more breaks than the previous year (due to change in operations). The Table below shows the validation results obtained with *machine learning*; the only model developed for this project.

ALL PIPES	% Breaks (2018-209) avoided (worst pipes first)		
% Pipes Targeted	**Desktop scoring**	**Multi-variable regression**	**Machine Learning**
1	NA	NA	**50.0**
5	NA	NA	**73.3**
10	NA	NA	**90.0**

info@infraplan-llc.com www.infraplan-llc.com

If the top **1%** worst pipes as ranked by their *machine learning* LOF had been replaced, **50.0%** of the 2018-2019 breaks would have been avoided; for the top **5%,** and **10%** worst pipes, the percentages are **73.3%,** and **90.0%**, respectively.

Project 3 illustrates the fact that even for a small system (less than 200 miles) with an erratic break pattern, machine learning yields excellent break prediction results; up to 90% of the breaks for the next two years are predicted.

How did we complete the predictions?

- **infraSOFT CLEAN module**. The *machine learning*-powered cleaning algorithm of CLEAN was used to clean the data for Projects #1 and #3. Data cleaning was more limited for Project #2.
- **infraSOFT STATS module** was used to detect data issues and interpret results.
- **infraSOFT PREDICT module.** The *machine learning* break prediction model of **PREDICT** was used to generate a LOF score, and forecast the future break number, for each pipe and at each year in the future.

Discover infraPLAN

infraPLAN has pioneered the use of advanced analytics in the field of water pipes R&R planning. The combination of our industry-leading platform infraSOFT and extensive consulting experience enables utilities and their consultants to generate data-driven R&R forecasting answers they can interpret and trust.

https://www.infraplan-llc.com

Annie Vanrenterghem Raven, PhD, CEO

Discover infraSOFT

Our platform, from two decades of field and research study, infraSOFT puts the power of Machine Learning at your fingertips to help you optimize the rehabilitation and replacement plan of your water pipes.

https://www.infraplan-llc.com/infrasoft

Discover our “Ask The Experts” articles

Where we address your questions about applying advanced analytics to water pipes R&R planning.

https://www.infraplan-llc.com/articles