Advances in Reliability, Safety and Security

ESREL 2024
Collection of Extended Abstracts

Reliability Analyses With Extreme Gradient Boosting And Shapley Additive Explanations

Florent Brissaud

Research & Innovation Center for Energy (RICE), GRTgaz, France

Keywords: reliability, machine learning, gradient boosting, shapley values

GRTgaz is a French gas transmission system operator (TSO). As a European leader in this sector, GRTgaz operates more than 32,500 kilometers of buried pipelines to transmit gas from suppliers to consumers connected to its network. The industrial assets of GRTgaz also include 26 compression stations and over ten thousand network substations (for sectioning, pigging, metering, pressure-reduction, and delivery) distributed across French territory. In recent years, the development of biomethane has also led to the addition of new installations on the gas network, such as injection units and small compression stations. GRTgaz then manages tens of thousands of industrial equipment items, including motorized and manual valves, safety closure and relief valves, gas regulators, gas analyzers, etc. to ensure the safe and efficient operation of the gas transmission network.

The industrial asset management policy of GRTgaz relies, in particular, on equipment databases and maintenance operation inventories. These data are collected through the Computerized Maintenance Management System (CMMS) since 2005. Equipment characteristics include details such as the position of the equipment within installations, its type, commissioning dates, calibration values, manufacturer references, flange diameters, flow capacity, and more. The maintenance inventories provide information on failure dates, as well as observed failure modes and causes. The processing of this equipment and failure data allows for the assessment of the health of industrial assets (based on criteria such as performance, safety, environmental impact, and costs). It enables continuous improvement in the management of supplies (both equipment and services), adaptation of maintenance policies, and prioritization of investments for renovation purposes. To this end, statistical reliability analyses have been conducted on these data since the early collection years (Brissaud et al., 2011, 2019). However, the fact that a certain amount of data is missing, erroneous, or inconsistent has prompted GRTgaz to explore more “flexible” analysis approaches, particularly those based on artificial intelligence techniques (Belounnas et al., 2022), as presented below.

Extreme Gradient Boosting (XGBoost) is a powerful machine learning algorithm that has gained immense popularity in both research and industry (Chen and Guestin, 2016). It falls under the category of ensemble methods, which combine multiple weak models to create a strong predictive model. XGBoost is particularly effective for regression and classification tasks. At its core, XGBoost is an extension of the traditional gradient boosting algorithm. It leverages a tree-based ensemble approach, where each tree is built sequentially to correct the errors made by the previous ones. XGBoost’s combination of regularization, gradient-based optimization, and parallelization makes it a robust and efficient choice for a wide range of machine learning tasks.

Shapley Additive Explanations (SHAP) is a technique used to explain the results of machine learning models (Su-In and Scott, 2017). It is based on Shapley values, which draw from game theory to attribute credit for a model’s prediction to each feature or feature values. To process, SHAP decomposes a model’s output into the sum of the impacts of each feature. It then calculates a value representing the contribution of each feature to the model’s result. These values help understand the importance of each feature and explain the model’s prediction, notably those provided by extreme gradient boosting.

This communication outlines the approach developed by GRTgaz for:

- preprocessing data coming from the CMMS: formatting (type conversion, unit harmonization, etc.), inferring (e.g. date and manufacturer corrections), and grouping (for sparsely represented values) categories, numerical data, and dates;
- calculating reliability indicators for analysis: determining the time to failure (TTF) for each line corresponding to a failure event, or the mean time between failures (MTBF) for each line corresponding to an equipment item, excluding failures considered as "repeated" and including right-censoring;
- analyzing reliability factors (or features): identifying the factors with the greatest impact on equipment reliability and defining equipment families with similar characteristics related to these factors.
- estimating reliability: determining failure rates as function on age for each defined equipment family.

All of these developments have been programmed in Python, and an interface has been created using pyQT to provide a tool that can be easily handled by the technical and operational teams. In addition, a direct connection has been established with the "data lake" of the CMMS data so that all these analyses are performed directly using the latest updated data.

Examples of results for an equipment item of gas pressure-reduction and delivery stations are presented in Figure 1. In Figure 1 (a), we observe that the most impactful factors on the reliability of this equipment are the year of commissioning of the item and of its station. However, the same graph reveals, through clustering (on the right), that these two factors are (of course) highly correlated. In Figure 2 (b), we see that as the year of commissioning decreases (indicating older equipment), the failure rate increases (as indicated by the higher SHAP values on the x-axis). Additionally, in Figure 1 (a), we notice that flow capacity is also an important factor, while being correlated with other factors such as manufacturer reference, position of the item, and flange characteristics. Pressure-related features are also significant, and, of course, they are correlated with each other. Finally, in Figure 1 (b), we observe that higher flow capacity or pressure (indicated by red points) corresponds to a higher failure rate.

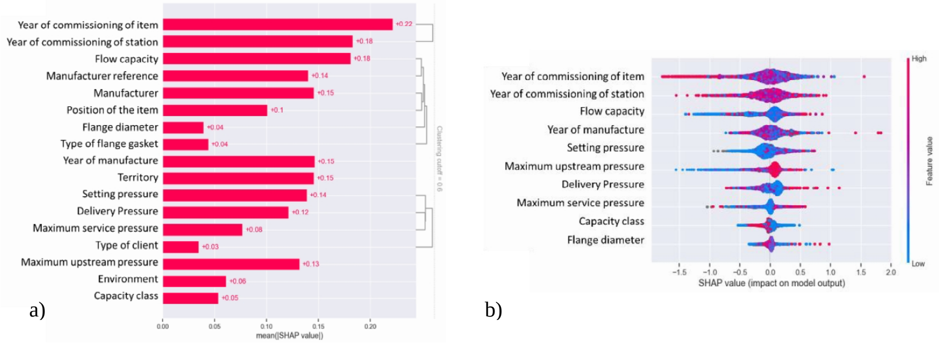

Fig. 1. (a) Reliability factors with degree of importance and correlation-based grouping; (b) Reliability factors with impact of values.

Other analyses, for instance, allow to identify which type of equipment (manufacturer references) are the most reliable based on operational conditions (flow capacity, pressure, etc.) in order to guide decisions according to the design of installations. Furthermore, the effects of age help target preventive maintenance operations for equipment most sensitive to aging. Lastly, configurations of equipment and operational conditions that pose the highest risk are prioritized in an industrial asset renovation program.

References

Belounnas, A., Brissaud, F., Rousset, E. 2022. Using Artificial Intelligence Algorithms to Identify Factors of Methane Leaks from Gas Transmission Assets. Proceedings of the 32nd European Safety and Reliability Conference.

Brissaud, F., Lanternier, B., Charpentier, D. 2011. Modelling failure rates according to time and influencing factors. International Journal of Reliability and Safety 5(2), 95-109.

Brissaud, F., Marle, L., Faure, D. 2019. Reliability Factors Analyses for Gas Transmission Items. Proceedings of the 29th European Safety and Reliability Conference.

Chen, T., Guestin, C. 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Su-In, L., Scott, L. 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30, 4765-4774.