21, rue d'Artois, F-75008 PARIS
http : //www.cigre.org

CIGRE US National Committee
2021 Grid of the Future Symposium

Using Machine Learning to Predict Ratio Transformer Failure

R. THAIKKAT[1], V. DECHIARO[2], J. GARRITY[1], S. MCCORMICK[1], R. POLERA[1], A. MACMULLEN[1]
Tagup, Inc.[1], National Grid[2]
USA

SUMMARY

Using data from across National Grid's upstate New York service territory, we constructed two models to accurately predict key performance outcomes for a fleet of network transformers. In order to improve productivity and reduce costs for National Grid's procurement department, we focused on two key performance outcomes in our model development: concordance index and remaining useful life error. After initial modeling, we addressed issues of miscalibrations due to the disproportional nature of the dataset by including three-phase transformers, imputing missing data, and adding a test dataset to the training and validation dataset. With these improvements, our new model is validated in a backtesting context and postured for forward testing. The efficacy of our approach is ultimately judged against the preliminary success criteria: overlap coefficient and missed preemptive removals.

KEYWORDS

Machine learning, failure prediction, time-to-event, distribution transformers, asset management

rthaikkat@tagup.io

Introduction

We have developed and deployed machine learning (ML) methods that quantify the likelihood of critical event occurrences. For electric utilities, critical events may include equipment faults or failure. By accurately estimating when these critical events occur, operators and engineers can make better asset management decisions. The result: improved equipment safety, reliability, and operating efficiency[1].

In partnership with National Grid, we are demonstrating the technical viability of deploying time-to-event[2] (TTE) models in a software application. Our software provides a prognostic tool for quantifying risk of ratio transformer failures across National Grid's upstate New York service territory.

[1] Operating efficiency refers to both equipment performance and the business processes (and associated costs) of managing a large fleet of assets like distribution transformers.

[2] TTE, also known as Time-to-Event modeling, is a machine learning-based technical approach for estimating event probabilities in a given period given any/all historical data preceding the start of that period (e.g., equipment static attributes, time series sensor measurements and event training data).

The majority of ratio transformer failures occur during summer when the system experiences peak loads due to increased use of indoor cooling systems. Prior to the summer season, an overloaded transformer list (henceforth, the "summer prep list") is prepared based on the expected peak loads. National Grid's Distribution Planning & Asset Management (DPAM)[3] group cross-references this list with additional information, such as location and age, to prioritize preemptive replacement and upgrade work for the upcoming season. The analytics described in this paper provide a holistic planning tool to augment the team's existing method by supplying model outputs informed by historical operating data and static attributes; volumes of data that may be unintuitive or difficult to systematically analyze. By improving this selection process, DPAM engineers can prioritize transformers with the highest probability of failure and reduce the chance of removing healthy assets.

National Grid's procurement department is responsible for purchasing replacement assets while maintaining optimal inventory levels by line item (or SKU). By identifying which transformers are expected to fail in an upcoming purchasing cycle (aggregated by purchasing code and location), the procurement team can make more informed purchasing decisions based on aggregated asset level failure risk. This in return allows National Grid to minimize business interruption from a shortage in replacement parts meanwhile reducing the cost of carrying excess inventory. Furthermore, by considering asset level failure risk geographically and managing inventory levels by location, engineers and resource planners can better identify whether inventory at a given storage location matches the needs of that service territory and assign the right size maintenance team required to support replacement and upgrade operations.

Model development efforts were focused on achieving two key performance outcomes: a high concordance index (CI) and low remaining useful life (RUL) error[4]. These goals must be balanced with proper model calibration, which ensures that the predicted event rates (i.e. how many transformers may fail) align with true event rates over a given time interval. It is well understood that there are trade-offs between a model's discriminative capabilities (as measured by CI) and the calibration of its predictions (measured by event count errors). This purported inversely proportional relationship means that a miscalibrated model may perform artificially well, whereas a well calibrated model may perform worse with respect to CI and RUL error.

We have iteratively reconfigured the model to address miscalibration. Broadly, miscalibration resulted from the disproportional nature of the dataset, which captured the complete set of installation records but only a very small fraction of observed removal records[5], consequently limiting the volume of transformer retirements with which to train the model. Miscalibration from this imbalance was compounded by narrow prediction horizons used for model evaluation and selection. When training a model to minimize RUL error, an observed failure from the validation set must have a RUL less than or equal to the prediction horizon. Thus, when the prediction horizon is very short, the model is

[3] DPAM, or Distribution Planning and Asset Management, is the target user organization within National Grid. DPAM is made up of over 50 engineers that are responsible for the summer prep planning process among other engineering and planning responsibilities.

[4] RUL error is the difference between the actual time of failure and the predicted time of failure.

[5] Asset lifetime (calculated based upon a known install and removal date) is a critical component to computing baseline hazard: an underpinning of the CLV survival model. The censored dataset (install base) was more complete than the uncensored dataset (removal records), which often could not be mapped to existing primary ID and always lacked an official removal date. We mapped retirements by primary ID to the install base to the maximum extent possible, inferring removal dates as needed. However, the inability to map retirements to existing primary ID resulted in a vast majority of retirement records being unusable, versus the majority of surviving transformers.

incentivized to minimize RUL errors, thereby over-predicting risk: an undesirable side effect. These factors culminate in a model that grossly *over predicts* asset failure (see Model 1 in Table 1).

Data transformations, among other techniques, may be used to scale the event counts to correct the over prediction issue. However, data transformations don't overcome the inherently disproportional nature of the dataset. Thus, this model manipulation tends to scale the event count by echoing the imbalanced ratio of retired assets to install base[6]. This phenomenon is shown in Model 2, Table 1.

We addressed this issue via random downsampling of the number of censored (installed) examples to match the same relative fraction of uncensored (removed) examples. This strategy resulted in a model that predicted an event count that more closely resembles reality, but modestly decreased the CI and substantially increased RUL error (see Model 3, Table 1). While this model is generally considered an improvement over previous iterations, we identified a series of improvements to be explored in the following section.

Table 1: A summary of various Model I model outputs. These results acutely demonstrate the nuanced relationship between various model calibration techniques and their impact on validation metrics and event counts.

Model	Description	Median RUL Error	CI
0	Baseline	7.3 years	0.54
1	Uncalibrated, Model I (single-phase)	1.6 years	0.87
2	Calibrated, Model I (single-phase)	1.6 years	0.87
3	Calibrated, Downsampled Model I (single-phase)	4.7 years	0.83

Model Improvements

These improvements broadly included the following:

- Including three-phase transformers, which had previously been excluded from previous models. The data imputation heuristic we employed to infer retirement dates for the observed three-phase removals was initially suspected to be problematic and thereby, possibly limiting the model's calibration capabilities. Hence, three-phase transformers were previously excluded from model training and evaluation. Upon further experimentation, it was determined that the root cause for model miscalibration issues was the disproportional nature of the dataset; model performance issues were not closely linked to the data imputation heuristic for three-phase removals. Hence, the new training dataset for the Model II included data from both single-phase and three-phase transformers.

[6] Without adjusting for the imbalance of censored vs. uncensored training examples, the model was only aware of 10% of all removals. As a result, the low count of predicted removals were consistent with the fraction of removed to installed transformers in the training set (an approximate 1:10 ratio).

- Imputing missing data such as size, manufacturer, district, etc. to expand the available datasets. This allowed the inclusion of approximately 4,000 additional transformers.

- Adding a test dataset in addition to the training and validation datasets to ensure that the model isn't biased by training on data for which it could potentially be validated. During our initial model development experiments, model performance metrics were evaluated using data that was also partially used for training. More specifically, this data includes a subset of transformers that were a part of the model training as well as the model evaluation phases. However, there also exists a well-defined timeline within the operational lifetimes/ages of those transformers that were excluded from model training and included only during model evaluation. We call this "out-of-time" model evaluation, which is a reasonable but an incomplete approach. A more rigorous approach is to fully exclude a subset of transformers from the model selection experiments and finally test the best performing model on that subset. We call this "out-of-time and out-of-sample" model evaluation. This type of model testing provides more reliable estimates of the model's predictive capabilities.

- Selecting an earlier conditioning date (January 2010 vs. January 2018) to expand the validation dataset. This change helps us avoid overfitting the model, whereby the model has incentive to predict unduly high risk due to the narrow predictive window (as described in previous sections). See Table 2.

Note that the model was downsampled to avoid pitfalls of having an imbalance amongst censored and uncensored assets. See Table 3 for details regarding downsampling.

Table 2: Model specifications for the revised model.

Specification	**Value**
Training Set	January 1952 - January 1, 2010
Conditioning Date	January 1, 2010
Validation Set	January 1, 2010 - June 1, 2021
Validation Metrics	CI (1 year), Median RUL Error

Table 3 illustrates the random downsampling of the censored dataset in order to more accurately reflect the ratio of actual annual removals from the install base. When considering the count of uncensored validation examples on an annualized basis (i.e. over the 11 year validation period), the ratio of removals to install base is roughly 1.5%: a close approximation of the true ratio of annual removals to the broader install base.

Table 3: A summary of the count of training and validation examples for the uncensored and censored datasets.

Count of Unique DeviceIDs	Training	Validation
Uncensored (Removals)	293	408
Censored (Installed)	1993	2306
Total	2286	2714

Model Results

The updated model (Model II) had a slightly higher RUL error and lower CI than its predecessor (Model I). As described in previous sections, model 4 has a more realistic event count (better calibration) resulting from a longer predictive horizon (earlier conditioning date) and an expanded validation set (access to ~10X more removal records based on imputation methods). Ultimately, a well calibrated model produces higher fidelity results; thus, the miniscule difference in RUL and CI are negligible.

Table 4: Experimental results resulting from various calibration methods throughout Model I and II development efforts.

Model	Description	Median RUL Error	CI
0	Baseline	7.3 years	0.54
1	Uncalibrated, Model I (single-phase)	1.6 years	0.87
2	Calibrated, Model I (single-phase)	1.6 years	0.87
3	Calibrated, Downsampled Model I (single-phase)	4.7 years	0.83
4	Calibrated, Downsampled Model II (single- and three-phase)	5.4 years	0.82

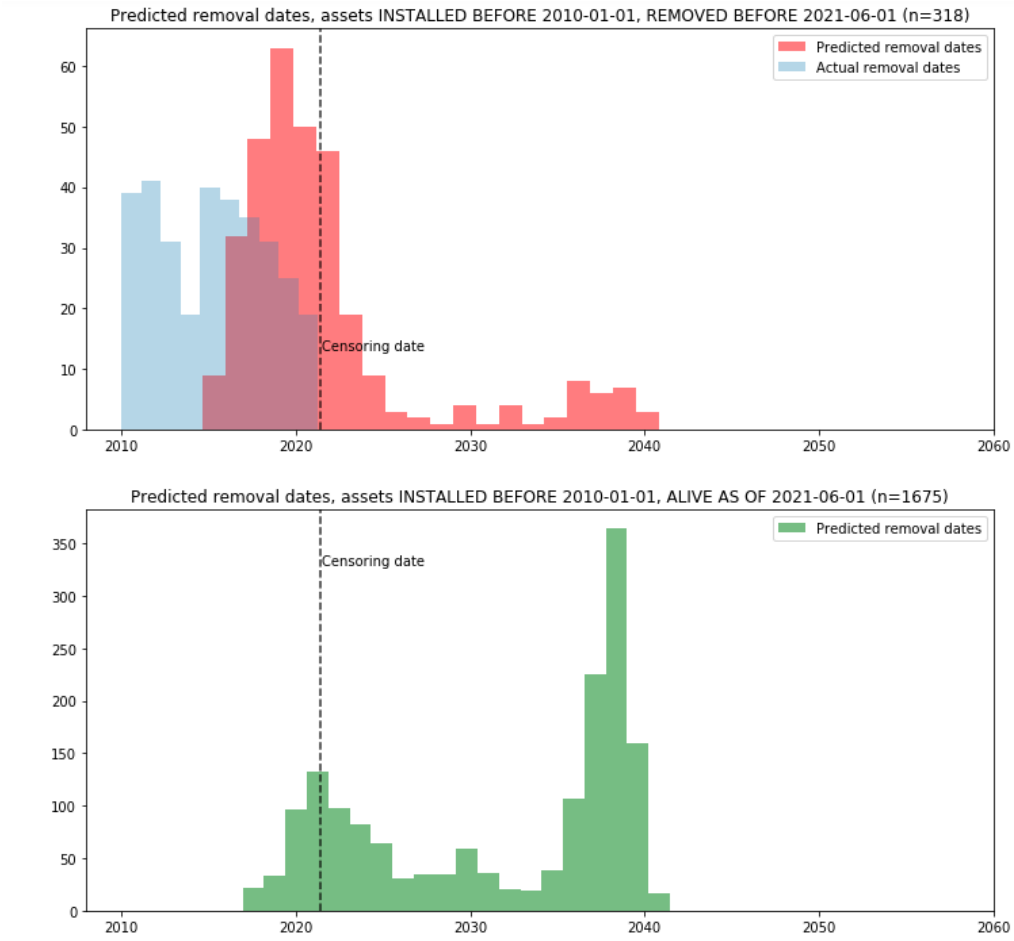
The discriminative power of the model is also evident in the below histograms in Figure 1. The upper plot demonstrates the distribution of predicted removal dates for all assets installed before the conditioning date (January 1, 2010) and removed prior to the censoring date (June 1, 2021). Conversely, the lower plot shows assets installed before the conditioning date that were still in place as of the censoring date. The two histograms collectively suggest the following:

- The model has reasonably good overlap between predicted removals and actual removals. The upper plot shows that a substantial portion of assets removed prior to the censoring date were predicted for removal by the model (indicated in the purple bins). Note that the blue bins

(distribution of actual removal dates) provide a visualization of the RUL error.

- The model can distinguish between high risk assets (low RUL) and low risk assets (high RUL). The clear right skew of the predicted removals (upper plot) compared with the severe left skew of the surviving assets suggests that the model is able to distinguish between assets that are expected to fail soon versus those that may survive well into the future.

Figure 1: Histograms showing the distribution of predicted vs. actual removal dates (upper plot) installed before the conditioning date (January 1, 2010) and removed before the censoring date (June 1, 2021) versus assets installed before the conditioning date that were still in place as of the censoring date (lower plot).



Model Validation

The above described model improvements culminated in back-testing validation of summer 2020 predictions and a rank ordered list of transformers to support forward testing validation at the close of summer 2021. In each case, we generated a "beeswarm" plot to compare the model's ranking of the following:

- Summer prep candidates. These assets are flagged for review by National Grid DPAM engineers based on loading estimates and other considerations.

- Summer prep preemptive removals. These assets were actually removed from service based upon engineering analysis; they represent a small fraction of the previously identified candidates.
- In-service failures. These assets failed in place. These failures are captured by the outage dataset.
- Remaining Install Base. These data points represent the balance of transformers that don't fall into the above categories.

The interpretation of these results and their contribution to preliminary success criteria is described below.

Preliminary Success Criteria

We submitted a rank ordered list of transformers by risk to National Grid on June 15, 2021. The forward testing validation period extends from submission to September 15, 2021. The efficacy of our approach will ultimately be judged against the preliminary success criteria[7]: overlap coefficient and missed preemptive removals.

The overlap coefficient measures the extent to which the model rankings overlap with the summer prep candidates list. Specifically, it is calculated by assessing how many of the *n* total summer prep candidates (e.g. n=1000) fall within the first *n* rank ordered model predictions, where perfect overlap would equal 1 and disjoint sets would return 0. This criteria serves two important purposes:

- **Provides an indication of how well our model captures peak loading signal.** The model does not use peak loading data as a feature due to our limited access to historical loading data[8]. However, the uncensored data (removals) fundamentally contains signal from the peak loading data given that peak loading percentages are currently the primary decision criteria for removal. In lieu of this data, the overlap coefficient provides an indication of how well our model picks up on the inherent signal contained in the training data.

- **Builds confidence with the end-user.** Model-based decision support can seem like a black box to new users which may discourage adoption. Higher overlap coefficients indicate that the model may, in part, "think like an engineer" by prioritizing similar transformers for removal. This familiarity, in addition to forthcoming feature importance capabilities integrated in the application, will assist in making model outputs relatable and interpretable to the end-user.

Missed preemptive removals correspond to in-service failures ranked highly by the model, but not captured in the summer prep list. This metric is the ultimate litmus test for how model-based decision support can augment existing workflows. For the purpose of the summer 2021 forward testing validation period, we have mutually agreed upon a target of *at least* one missed preemptive removal occurring at or below the 10th percentile[9].

---

[7] While these success criteria have been vetted and mutually agreed upon with National Grid, they may be subject to change as we continue workshops with the end user.

[8] We currently only have access to peak loading data from the TLM system for 2019 and 2020 for a subset of transformers.

[9] This is an undoubtedly challenging task, as in-service failures represent less than 0.3% of the install base, annually. However, a correct prediction (or even a near-miss) would present a compelling case for the value of model-based decision support.

While both of these metrics are reported as discrete numbers, they can also be visualized in the beeswarm plot. Ideally, this would look like **in-service failures** and **summer prep candidates** being to the very far left of the plot, indicating that the model ranked these assets as high risk. Next, we'll assess how our revised model performed against the 2020 summer prep list and set the scene for forward testing for the summer 2021 cooling season.

Summer 2020 Retrospective Evaluation

In summer 2020, National Grid DPAM engineers identified 435 candidates (indicated in green) for preemptive removal based upon peak load percentages exceeding 100% for the previous period, amongst other considerations. As shown in Table 6, nearly 40% of these candidates and removals were ranked in the top quartile based upon risk.

Figure 2: Comparison of National Grid summer prep list, preemptive removals, and in-service failures vs. model rankings for summer 2020.

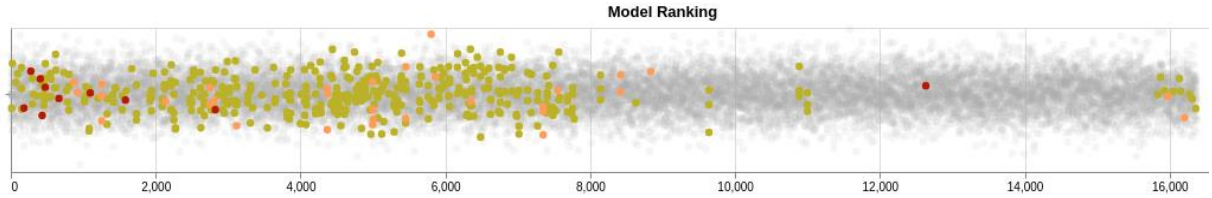


Table 5: Analysis of the 2020 summer prep candidates and preemptive removals (435 total) captured by the model in the top quartile.

	% Captured	# Captured	Percentile Value
1st percentile	1.6%	7	164
5th percentile	7.8%	34	819
10th percentile	13.0%	57	1638
25th percentile	39.5%	173	4095

Table 6: Analysis of the 2020 in-service failures (10/98 total) captured by the model in the top quartile. The remaining 88 outages could not be mapped to the install base at the time based upon missing data (i.e. device ID or primary ID). National Grid is working on re-extracting the data with these critical identifiers so we may complete our analysis of model performance on the 2020 summer prep program.

	% Captured	# Captured	Percentile Value
1st percentile	0%	0	164
5th percentile	60%	6	819

10th percentile	80%	8	1638
25th percentile	90%	9	4095

Figure 3: Comparison of National Grid summer prep list vs. model rankings for summer 2021.

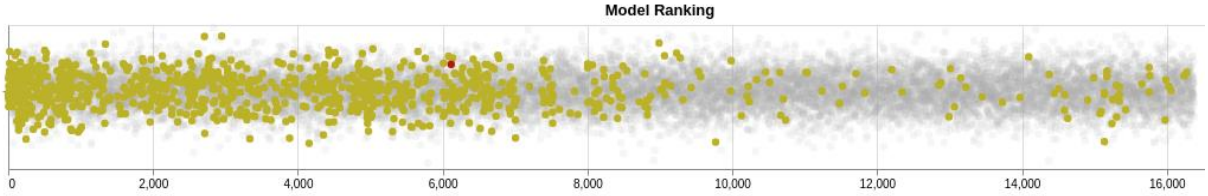


Table 7: A summary of the known inputs for the summer 2021 season. Summer prep preemptive removals and in-service failure counts will not be known in their entirety until the close of the validation period (September 15, 2021).

	Count
Summer prep candidates	979
Summer prep preemptive removals	TBD
In-service failures	TBD
Total install base	16336

Table 8: Analysis of the 2021 summer prep candidates (979 total) captured by the model in the top quartile.

	% Captured	**# Captured**	**Percentile Value**
1st percentile	6.7%	65	164
5th percentile	20.2%	197	819
10th percentile	28.5%	278	1638
25th percentile	54.1%	527	4095

Conclusion

Early models suffered from decreased model calibration due to a fundamentally disproportional dataset which captured the complete set of installation records but only a very small fraction of observed removal records. This phenomena was partially mitigated via downsampling (to reflect the true ratio of installed to removed records), but still resulted in a miscalibrated prediction of annual transformer

removals. The new model improved the calibration by expanding the prediction horizon and relatedly, increasing examples in the validation dataset. While the model had very minimal decreases in other performance metrics (i.e. RUL error and CI), the substantial increase in model calibration means that predictions made by this model are higher fidelity than any of its predecessors.

Next, the new model was validated in a back-testing context (2020 summer prep program) and postured for forward testing on 2021 summer prep at the close of the season. Both models are measured against mutually agreed upon preliminary success criteria. Preliminary metrics include the overlap coefficient and missed preemptive removals. The former generally measures the overlap between the summer prep list and model rankings, whereas the latter assesses how many in-service failures the model predicts (specifically at or below the 10th percentile) that were missed by the summer prep program. In summary, the 2020 summer prep program had a 3% overlap coefficient and captured 8 missed preemptive removals within the 10th percentile (80% of the total usable failure records). While we cannot yet evaluate missed preemptive removals for the 2021 summer prep program, the 24% overlap coefficient is a promising precursor to the hopeful capture of in-service failures: a result that could quantify the value creation potential of predictive analytics for transformer asset management.

BIBLIOGRAPHY

[1] McCormick, S. et al. Using Machine Learning to Quantify the Impact of Weather on Transformer Failure Risk. *CIGRE-US NGN paper competition, 2019* (unpublished).
[2] McCormick, S. et al. Using Machine Learning to Quantify the Impact of Heterogeneous Data on Transformer Failure Risk. *CIGRE-US Grid of the Future Symposium*, 2019 (unpublished).