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

CIGRE US National Committee
2019 Grid of the Future Symposium

Using Machine Learning to Quantify the Impact of Heterogeneous Data on Transformer Failure Risk

S. MCCORMICK, S. SHU, B. COOLEY, W. VEGA-BROWN, J. GARRITY
Tagup, Inc.
USA

B. KITTRELL
Con Edison
USA

SUMMARY

Using large-scale distributed computing and a variety of heterogeneous data sources including real-time sensor measurements, dissolved gas measurements, and localized historical weather, we construct a predictive model that allows us to accurately predict remaining useful life and failure probabilities for a fleet of network transformers. Our model is robust to highly variable data types, including both static and dynamic data, sparse and dense time series, and measurements of internal and external processes (such as weather). By comparing the predictive performance of models built on different combinations of these data sources, we can quantify the marginal benefit of including each additional data source in our model.

In order to relate each type of data to the risk of failure across a fleet of transformers, we have developed a novel class of survival models, the convex latent variable (CLV) model. This type of specialized survival model has several advantages. Rather than an opaque and subjective "health index", it produces interpretable predictions like the probability of failure within a given time window or the expected RUL of an asset. Our framework supports accurate estimates of the risk of equipment failure across a wide range of time-scales, from a few weeks to many years in the future, and can model not just the instantaneous risk of failure due to an event like a storm, but also the long-term impact on the risk of failure.

Our model has the added advantage of being highly interpretable compared with the vast majority of state-of-the-art machine learning models, many of which are based on deep learning approaches. The essentially linear structure of the CLV model allows the extraction of feature importances after training, enabling us to quantify the contribution of each measurement to the predicted risk of failure. We demonstrate that the feature importances learned by our model align with intuition and can, for example, be used to help asset fleet managers determine what types of maintenance actions are most beneficial for the overall health of their assets.

We empirically demonstrate that incorporating additional data sources increases the predictive accuracy of our models, as verified by validating predictions such as the number of transformer failures by line item, the primary categorization for purchasing decisions used by Con Edison. At a six-month predictive

jon@tagup.io

horizon, the average absolute error between predicted and actual failure counts for each line item category was less than one asset; at a one-year predictive horizon, the mean absolute prediction error was less than two assets.

This work complements our 2019 CIGRE-NGN paper submission [1] and serves as a continuation of our collaboration with Con Edison. This case study suggests these kinds of predictions could be used to improve financial planning for future capital expenditures or to justify investment in preventative maintenance or risk mitigation. Our approach is highly flexible and can easily be extended to incorporate other types of available data.

KEYWORDS

Asset management, predictive maintenance, survival modeling, dissolved gas, work requests, network transformers, failure prediction, interpretable machine learning

PROBLEM STATEMENT

Transformer failures are expensive, in terms of downtime, manpower, and equipment cost. If fleet operators better understood when particular transformers were likely to fail, they could substantially mitigate these costs with preemptive maintenance and removal of high-risk units. In addition, failure risk could be used to estimate key financial metrics in a principled and interpretable manner, allowing improved inventory management and procurement processes for the fleet. However, predicting transformer failure is challenging due to the complex array of internal and external factors that contribute to an asset being removed from service. We have previously demonstrated that machine learning and real-time sensor data can be used to assess failure risk, but remote sensing systems are expensive, and may be insufficient on their own to enable predictions that are accurate enough to be practically useful. To overcome this problem, we have developed a model that incorporates heterogeneous data sources and is robust to high rates of missing data. We demonstrate our model's ability to accurately predict transformer failures by backtesting on historical failure records.

DATA

Con Edison operates a fleet of approximately 28,000 network transformers in the greater New York City area covering the boroughs of Manhattan, Brooklyn, the Bronx, Queens, and the county of Westchester. A subset of these transformers is equipped with remote monitoring system (RMS) sensors, which enable real-time collection of information from each transformer. For this study, we restrict our attention to 4,664 of these transformers with intact RMS data. Our dataset contains information on transformer removals and RMS data from the beginning of 2009 until June 20, 2019, the last date on which data was collected.

For additional information on the Con Edison network transformer fleet, please refer to our CIGRE-NGN submission [1].

Static Data

The simplest type of asset-specific information we incorporate in our predictive model is static data, innate features of each transformer that do not change over time. These features include transformer rating, vault style, manufacturer, primary and secondary voltage, and cooling medium.

RMS Data

The remote monitoring sensors installed on the Con Edison transformers capture load, voltage, temperature, pressure, and six binary "flag" variables signaling dangerous conditions such as high oil temperature, water detected in the vault, and backfeed. These measurements are recorded roughly every 10 minutes.

Weather Data

We obtained sixty years of historical weather data via API from Dark Sky, a leading provider of localized weather data. The data is recorded at an hourly frequency from each of the New York City boroughs of Manhattan, Brooklyn, the Bronx, Queens, and the county of Westchester.

For more details on each of these three data sources, please refer to our CIGRE-NGN submission [1].

Work Order Data (WMS)

All field service data (e.g. work orders/requests) is stored in Con Edison's Work Management System (WMS). This dataset captures all field work related to preventative maintenance (e.g. CINDE time-based inspections, etc.) and corrective maintenance (e.g. PTO switch check, sump pump failures/replacements,

transformer replacements, etc.). Two of the nine features are free form text which include work request name and description.

The amount of WMS data is limited by the fact that valid records are only available from 2015 onwards, resulting in a total of 9,563 individual records. 3,201 of the 4,664 (68.6%) of transformers in our sample had at least one associated WMS record.

The following binary features were extracted from each raw WMS record:

- The category of the work order, determined by the priority code variable found in the raw data. This fell into one of eight categories: PTO switch check (low/medium/high priority), CINDE (remote/low/medium/high priority), and Other. Each record falls into exactly one of these mutually exclusive categories.}
- Whether the text description of the record contained the words “sump” or “flush”. We had initially experimented with including features denoting the presence of several other commonly found words, such as “replace”, “install”, and “relay”, but after an examination of the feature importances learned by the model these features did not appear to contain much useful signal and were therefore excluded from the final model.

Dissolved Gas Data (DGA)

The DGA data contains records of dissolved gas concentrations in oil (e.g. dissolved hydrogen, carbon monoxide, acetylene, etc.). When observed dissolved gas levels exceed certain thresholds the frequency of sampling may be increased. If dissolved gas levels exceed critical threshold levels, the transformer can be removed from service due to an increased risk of in-service failure. A total of 8,017 individual records were successfully mapped to our install base, resulting in 3,341 of the 4,664 (71.6%) of transformers having at least one associated DGA record.

Con Edison provided us with the dissolved gas thresholds they used for placing transformers into one of four categories that determines sampling schedule: “normal”, “add to watch list”, “resample every 12 months”, and “engineering evaluation CFR/OOE-2” - of which engineering evaluation typically calls for immediate removal from service.

As with the WMS data, we included a large group of possible features in our initial model, and by running experiments we determined which features were likely to be useful and which were not. Our final model included binary features for each of the following:

- A dissolved gas sample being taken
- Watch list/Resample/Removal status based on thresholds for each of the following gases: hydrogen (H_2), methane (CH_4), ethane (C_2H_6), and acetylene (C_2H_2)
- Levels of hydrogen, carbon monoxide, ethane, methane, and total combustible gas

METHODS

For a thorough treatment of the CLV model specification, please refer to our 2019 CIGRE-NGN submission [1].

Model Validation

To evaluate the predictive capacity of our model, we use backtesting, a process in which we train the model using only information available before a fixed date in the past---the conditioning date---then make predictions about the time between the conditioning date and the present, and use the withheld data to evaluate those predictions. Those predictions are validated using the metrics described below.

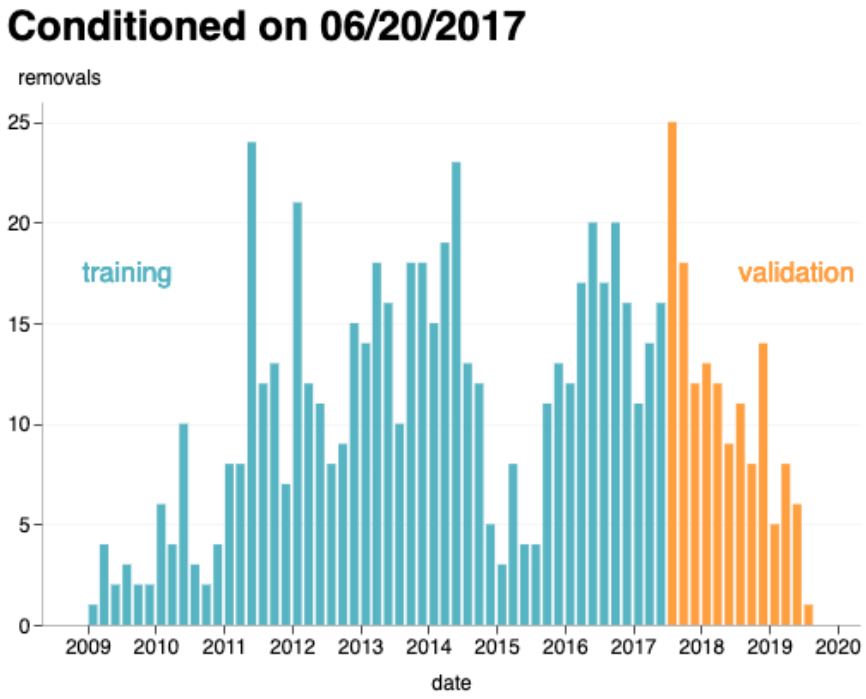

Figure 1: Histogram of transformer failures, partitioned into training and validation sets using the conditioning date June 20, 2017. Selecting this conditioning date enabled us to validate model predictions on the most recent two years of data. All experimental results reported were produced using this conditioning date.

Remaining Useful Life Prediction Error Statistics

We can use our survival models to estimate the remaining useful life (RUL) for any given asset. By comparing the predicted RUL to the actual time between prediction and failure, we can assess the accuracy of these RUL predictions for any transformers that failed after the conditioning date. We report both the mean and the median absolute RUL prediction error.

Concordance Index

The concordance index[1] (CI) measures how often the model correctly orders the lifetimes of the assets. A concordance index of 0.5 means that our model predictions are only as good as random guessing, while a concordance index of 1 indicates our model ordered all pairs of lifetimes perfectly. This metric tells us how well the model can rank the relative risk of the transformers but says nothing about the absolute accuracy of the RUL predictions.

Failure Count Predictions by Line Item

Con Edison purchases transformers by "line item", an identifier analogous to a SKU of a retail product. An accurate forecast of transformer failures by line item allows the procurement team to optimize purchasing and inventory management processes by:

- Accurately allocating capital on a year over year budget cycle
- Minimizing excess inventory with just-in-time (JIT) inventory management
- Negotiating bulk ordering discounts by minimizing 1-off, spot purchases/replacements

[1] The concordance index is equivalent to the area under the ROC curve (ROC-AUC) metric commonly used for measuring the performance of a binary classifier.

The predictions of the CLV model pertaining to individual transformers can be aggregated by subcategory, allowing us to predict failure counts by line item within any time horizon we choose. At each of these horizons, we compute the mean absolute error (MAE) between predicted and actual failures in this horizon. We consider only the 10 line item categories with the most failures in this calculation, as the number of installed transformers within each line item category drops off significantly after the top 7 categories.

Failure Probability Calibration Curves

A failure probability calibration curve compares predicted versus actual failure rates for transformers grouped by *predicted* failure risk. The simple example in Figure 2 (generated from actual one-year failure probability predictions output by the CLV model on Con Edison data) illustrates the relationship between predicted failure rates and failure counts by group. The diagonal line running through the right subplot represents the line of perfect accuracy, where the average predicted failure rate within a group of transformers exactly matches the actual failure rate for that group. The marker size of the prediction indicates the number of failure probability predictions that fall into that group. Although we generally group our predictions into 15 or 20 bins to get a more precise estimate of calibration quality, for instructive purposes we group the transformers into three categories in Figure 2: low, medium, and high risk. The corresponding bar chart in the left subplot shows how the distance of the dot in the right subplot from the optimal prediction line corresponds to the error in the predicted count of failures for that group of units.

When comparing validation metrics between different models, we report the weighted mean square error (WMSE) between the average predicted failure rate in each group and the actual failure rate. We weight the simple squared error by the number of predictions in each group, since a well-calibrated model should more closely approximate the actual failure rate as the number of units in each group increases.

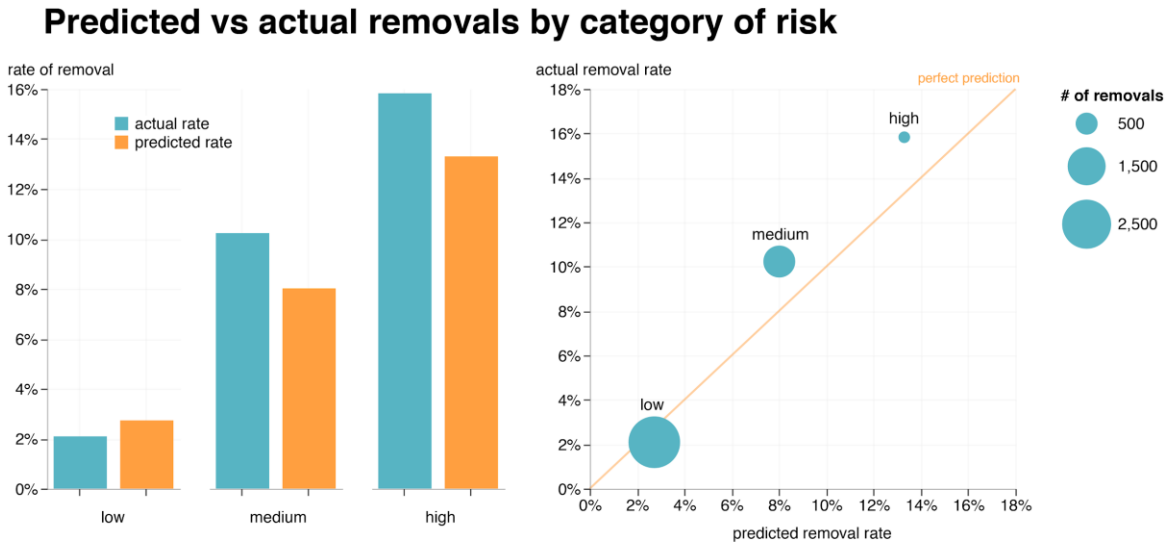

Figure 2: The failure probability calibration curve reflects both a model's ability to differentiate between high- and low-risk transformers and make accurate predictions of failure counts.

RESULTS

DGA/WMS Feature Importances

To better understand the effects of WMS and DGA data on transformer longevity, we trained a model using static, WMS, and DGA data on all available data (up until June 20, 2019, the last date of data collection). Feature importance is extracted from the parameter values learned by the model that encode

the relationship between each feature and expected failure risk. Because the features fed into the model are centered and scaled (meaning they have roughly mean zero and a standard deviation of one), a positive feature importance indicates that the model has learned that a feature correlates positively with failure probability. Conversely, a negative feature importance implies that that feature is inversely proportional to failure probability. The feature importances learned by the model and their relative magnitudes are illustrated in Figure 3.

Nearly all of the DGA features appear to be positively correlated with failure probability, which confirms our original hypothesis; higher concentrations of the dissolved gases monitored by Con Edison indicate the presence of electrical and thermal stresses within the transformer, making imminent failure more likely. The most important feature by far is a binary variable simply indicating whether or not a sample was taken. Occurrences where watch thresholds for hydrogen and methane are exceeded also appear to be particularly important, and the most important continuous DGA feature appears to be the level of total combustible gas in the sample.

The learned importances for the WMS features are less significant in terms of magnitude than for the DGA data. First, the PTO switch check feature does not provide much signal related to failure; the model has determined that all three binary features corresponding to “low”, “medium”, and “high” PTO work requests contribute very little to predicted failure risk. The two features extracted from the raw text records (the presence of the words “sump” and “flush”) appear to correlate positively with the remaining useful life of an asset, possibly because they indicate maintenance actions to fix problems that would have otherwise resulted in failure soon afterwards. CINDE inspections are part of Con Edison’s routine preventative maintenance program, and therefore the fact that the model has learned that CINDE inspections (in particular, high priority CINDEs) lead to longer transformer lifetimes on average seems reasonable.

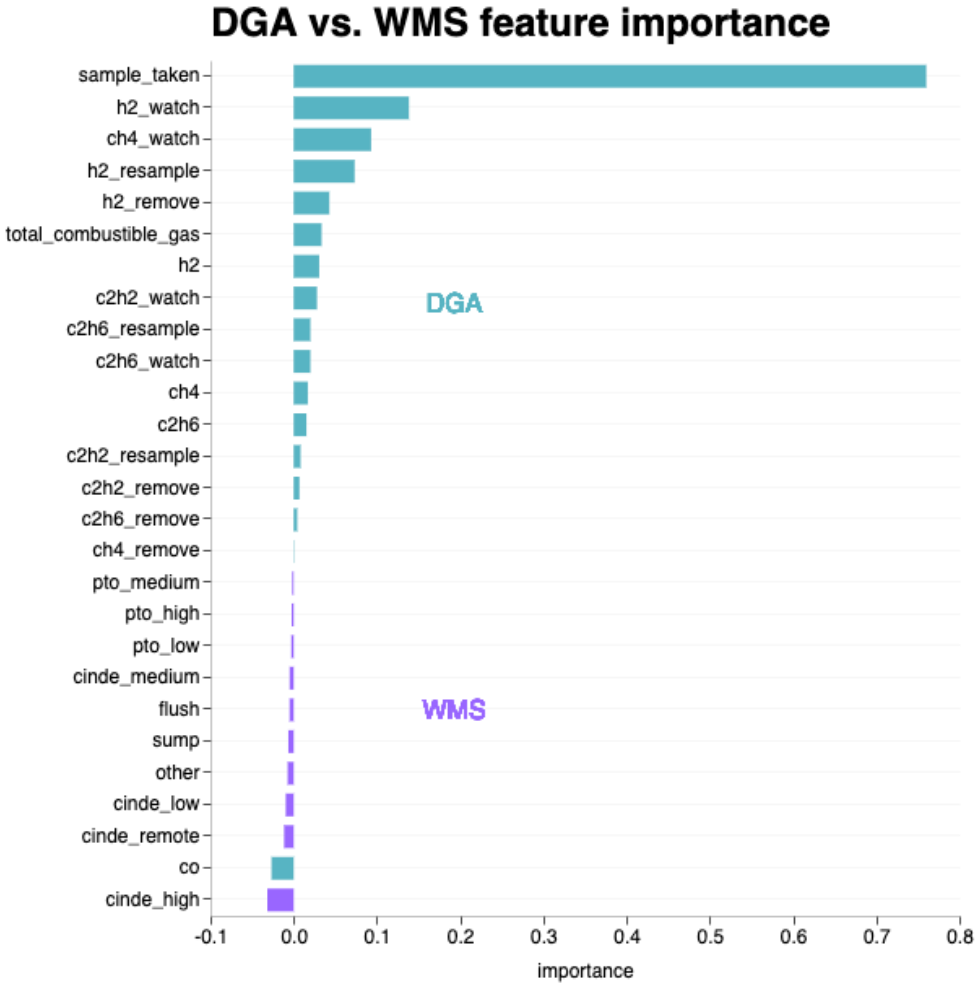

Figure 3: Learned feature importances for a CLV model trained using static, DGA, and WMS data.

The feature importances in the figure above suggests that DGA data contains more predictive signal than WMS data. In the following section, we run a series of experiments using different combinations of data sources in order to determine the marginal impact of incorporating each data set on the predictive capacity of the CLV model.

DGA/WMS Predictive Performance

The feature importance results indicate that the DGA data contains some signal that will allow the model to better predict failures. However, it's entirely possible that the same signal already exists in the RMS data, meaning that adding the DGA data to the static + RMS + weather model might not improve predictive performance. We tested this hypothesis using a conditioning date of 06/20/2017; the results are reported in Table 1.

Models compared:

(1) Static only
(2) Static + LIMS
(3) Static + RMS + weather (Task 9 model)
(4) Model 3 + LIMS
(5) Model 3 + LIMS + WMS

Model	CI (3mo)	CI (6mo)	Calibration WMSE (1yr)	Calibration WMSE (2yr)	Line item MAE (6mo)	Line item MAE (1yr)	Line item MAE (2yr)
(1)	0.73	0.78	0.07	0.18	0.91	1.7	5.67
(2)	0.75	0.79	0.05	0.17	0.91	1.77	5.8
(3)	0.77	0.81	0.05	0.08	0.99	1.67	4.92
(4)	0.77	0.81	0.04	0.1	0.99	1.68	4.92
(5)	0.77	0.81	0.04	0.1	0.99	1.67	4.91

Table 1: Validation scores for each of the five CLV model specifications.

By comparing the validation results of Models 1 and 2, we see that adding DGA data to a static-only model improves predictive performance of the CLV model. However, when considering the performance of Models 3 and 4, there appears to be very little increase in predictive accuracy when incorporating the DGA data into a model built using RMS and weather data. This is not particularly surprising when considering the relative volumes of RMS and DGA data available: the average number of individual RMS observations for each transformer in our sample was 128,134, while the average number of DGA observations per transformer in our sample was just 1.72. Of course, if the signal allowing us to predict failure in the RMS data and the signal in the DGA data were completely uncorrelated, the model should still improve when adding a new data source containing novel information. However, because the information contained within both data sources can reflect internal faults within the transformer, it's reasonable to imagine that the majority of the signal contained within the DGA data can also be drawn out of the RMS data. This explanation would corroborate our experimental findings: adding DGA data to a static-only model improves performance but adding DGA data to a model with RMS and weather data does not result in higher predictive accuracy.

Validation Results for the Best Performing Model

We present visualizations of the validation output for the best-performing model built using all data sources (Model 5). The model was trained on all data before June 20, 2017 and the predictions were compared to the most recently recorded two years of failure data (06/20/2017-06/20/2019).

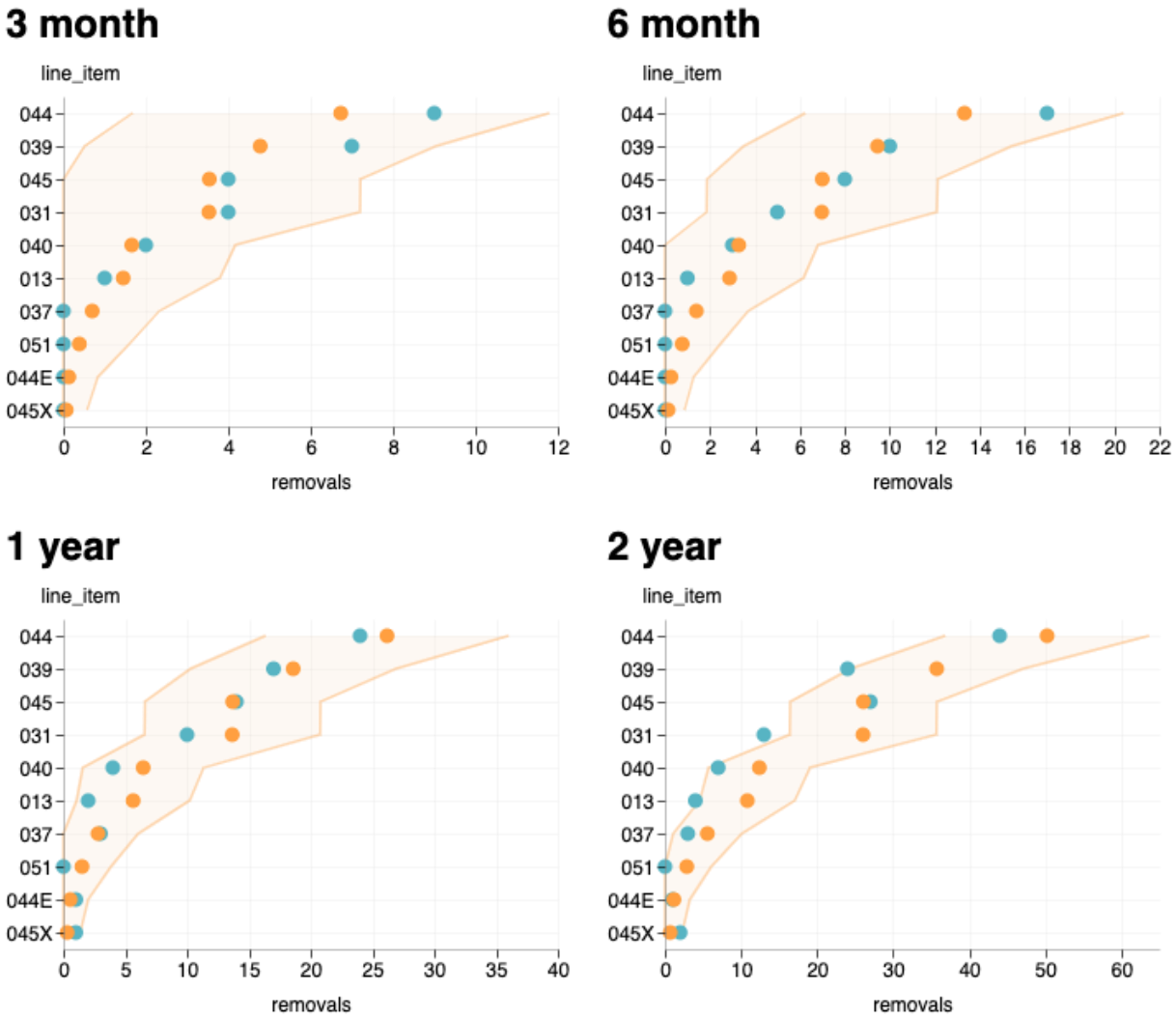

Figure 4: Failure predictions by line item category validated on the most recently observed two years of failure data from the Con Edison fleet.

Figure 4 shows predictions of failure counts by line item at each of the four considered time horizons, for the top ten most commonly installed line item categories. The predictions are very accurate with the exception of the two-year predictions, at which the model consistently overpredicts failures. This is likely due to a significant drop in the number of failures in 2019, possibly because recent failures had not yet been logged in the version of the data we receive from Con Edison.

Figure 5 shows failure probability calibration curves for the best performing model. At all four considered horizons, the predicted failure rates tend to closely match the actual failure rates when grouping by predicted failure risk. Since failures are rare, the model predicts fewer transformers to be at a high risk of failure, and the size of each group steadily declines as the predicted failure rate increases. As a result, the statistical uncertainty (variance) of our predictions increases for those high-risk groups, evidenced by the high variability of the dots farther to the right in each subplot. We quantify this uncertainty by computing 95% confidence intervals for our failure rate prediction, depicted as the blue shading in the plot.

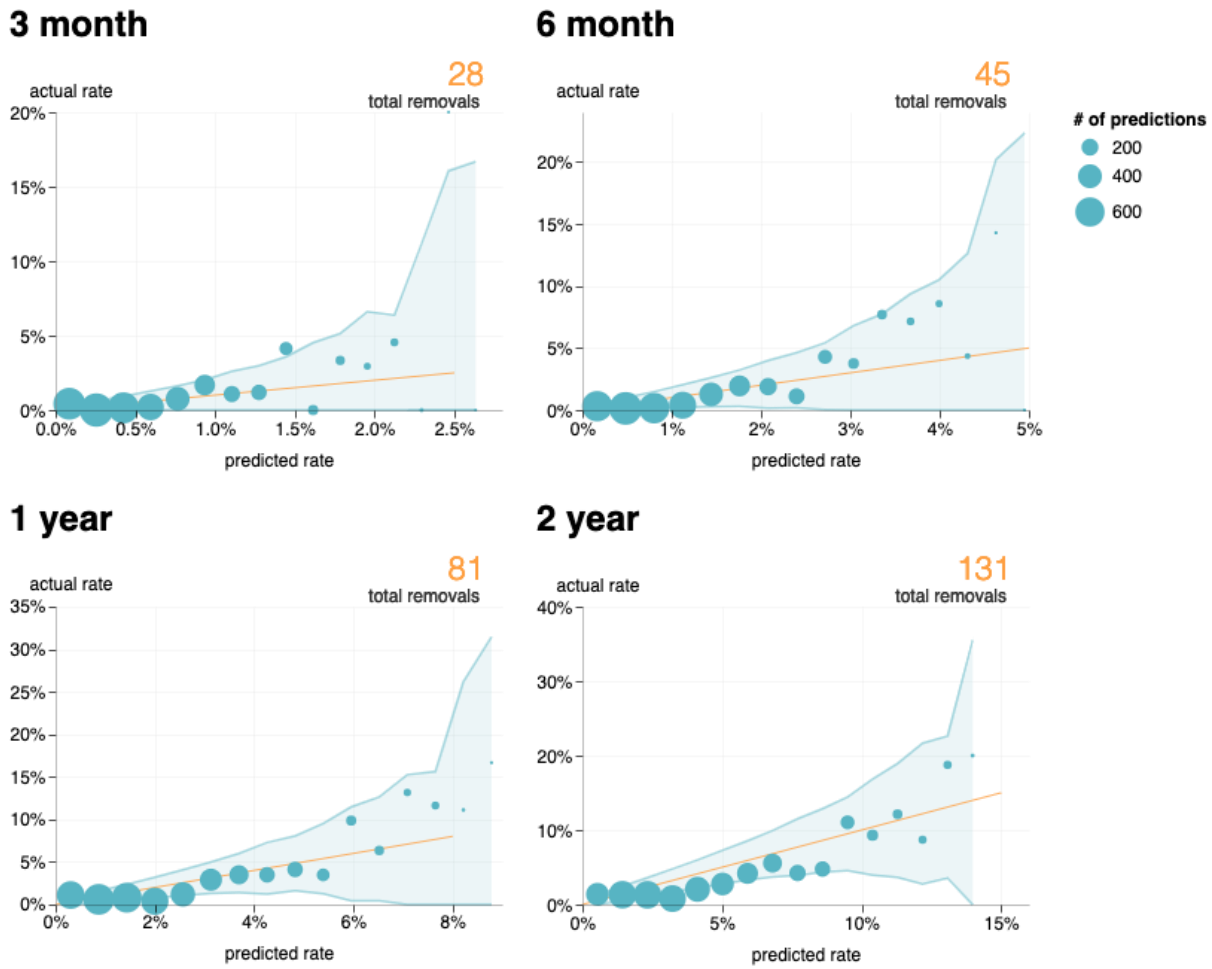

Figure 5: Failure probability calibration curves validated on the most recently observed two years of failure data from the Con Edison fleet.

CONCLUSION

By exploiting the flexibility of the CLV model, we used five distinct data sources to make predictions of failure risk for a large subset of Con Edison's network transformer fleet. We used the feature importances extracted from the trained model to quantify and compare the relative effects of DGA and WMS features on failure probability. We found that adding DGA to a model built only on static data improved predictive performance but incorporating DGA in a model built on RMS data caused no noticeable change in our validation metrics.

We set up an experiment in which we trained our model on historical data observed before June 20, 2017 and predicted transformer failure counts across the fleet between 6/20/2017-6/20/2019. At a six-month predictive horizon, the average absolute error between predicted and actual removal counts for each line item category was less than one asset; at a one-year predictive horizon, the mean absolute prediction error was less than two assets. In practice, these predictions could be used to optimize the allocation of capital for replacement transformers and minimize excess inventory.

In future work, we hope to expand our model to novel data sources and improve upon our existing feature engineering methods; for example, using known analytic approaches such as Duval pentagon to extract useful information from DGA data. As we gain a better understanding of the key factors underlying machine failure, we expect both the interpretability and the predictive utility of our model to increase.

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).