National Grid

Addressing Compliance Risk with Machine Learning

UA Week

November 1, 2023

nationalgrid

Agenda

01

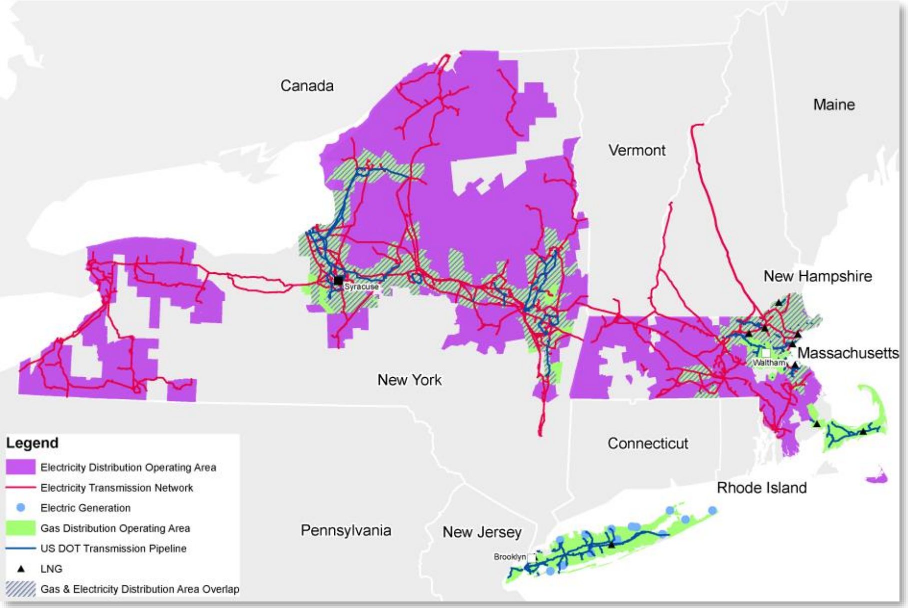
Who we are

National Grid

nationalgrid

About National Grid

We are one of the largest investor-owned energy companies in the US — serving more than 20 million people throughout New York and Massachusetts.



Serving 20 million people

5.3M Residential + 600k Commercial

= 5.9 million customer accounts

Residential & Commercial customers by region:

UNY 1.7 million
LI 0.6 million
NYC 1.3 million

nationalgrid

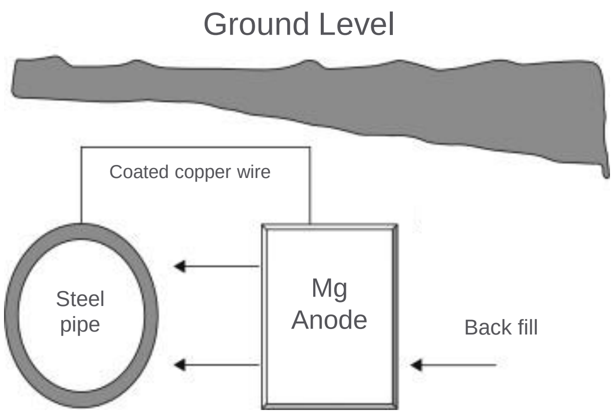
02

Cathodic Pipe Protection

Guaranteed Streets

nationalgrid

Cathodic Pipe Protection (CPP) – a brief primer



https://www.sciencedirect.com/topics/engineering/sacrificial-anode

Basic Idea:

- Cathodic Protection is a method of keeping steel from corroding
- An electrical connection in moist soil to a buried "sacrificial anode" moves electrons allowing the anode to be sacrificed and suffer the damage instead of the pipe

What do we need to know for this discussion?

- National Grid has thousands of test points along its network of gas pipelines in Massachusetts to monitor the voltage difference between Pipe and Soil
- When a test point reading, most inspected annually, has a difference less than -0.85 Volts, **intervention must be taken**

Business Case: Getting ahead of CPP compliance Issues

Risk: Newly paved roads in Massachusetts are under “guarantee” for 5 years, which prohibits digging.

- National Grid **cannot address CPP compliance issues** on guaranteed streets until the guarantee period is over.

Solution: Use machine learning to estimate the **probability that test points will have a failed inspection** during the compliance period

- Allows National Grid to react proactively to paving notices
- Additional Use Case: Inform CPP maintenance prioritization

03

Corrosion Data

Exploration and Initial Modelling Attempts

nationalgrid

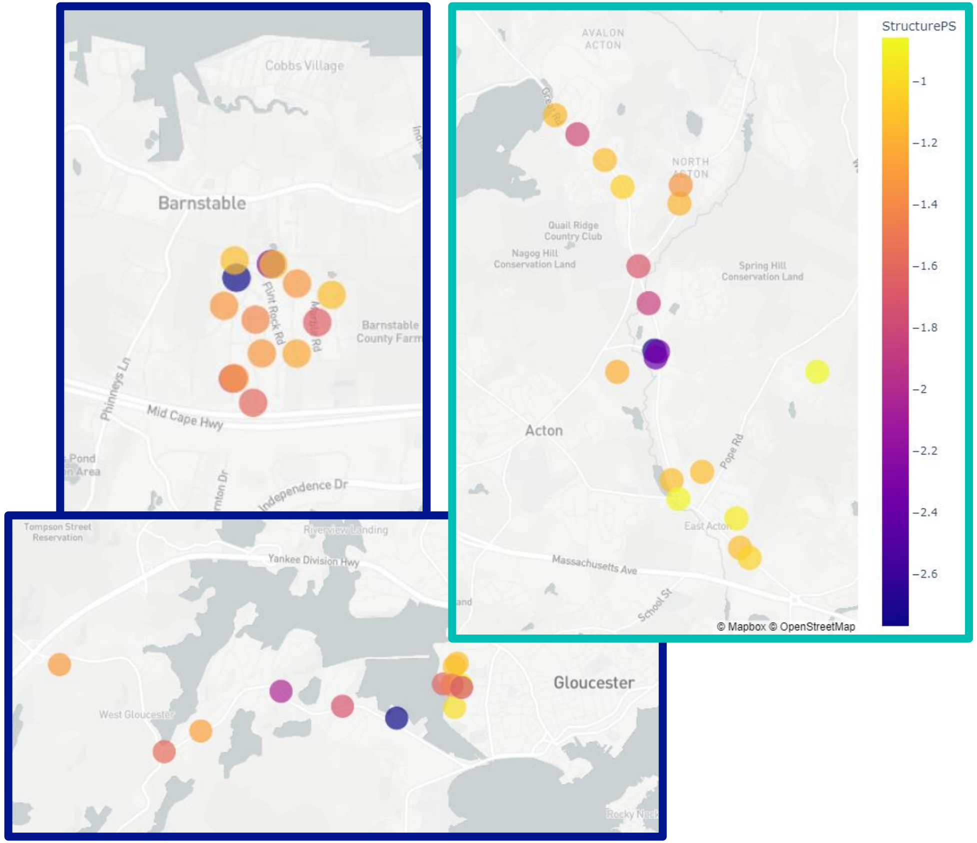
- **Corrosion Data**
  - Summary and basics of the structure
  - First look at our critical variable
  - Other key variables
- **Model Graveyard**
  - Original Plans
  - Regression & Survival Analysis
  - Pivot

- **Corrosion Data**
  - **Summary and basics of the structure**
  - First look at our critical variable
  - Other key variables
- **Model Graveyard**
  - Original Plans
  - Regression & Survival Analysis
  - Pivot

A summary of the data

- Our dataset spans from January 2000 through September 2022
- Nearly **30,000 Inspections** are carried out **each year**
- Each test point has information about its **location**, the **section** of pipe to which it belongs, **maintenance** history, inspection **notes**, and many more variables
- Our critical data point is **Pipe-to-Soil voltage** reading (recall -0.85 threshold)

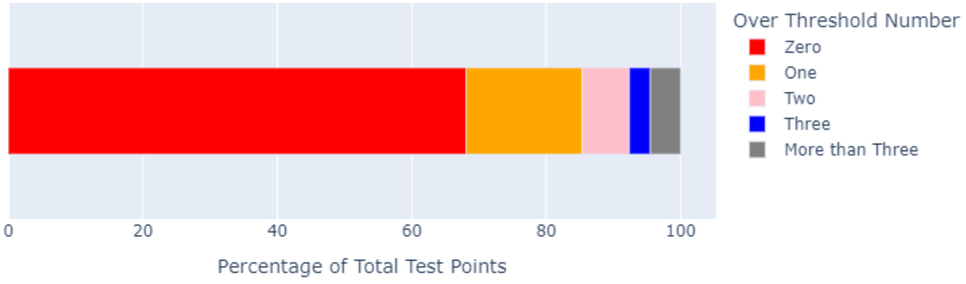
Pipeline sections and test points



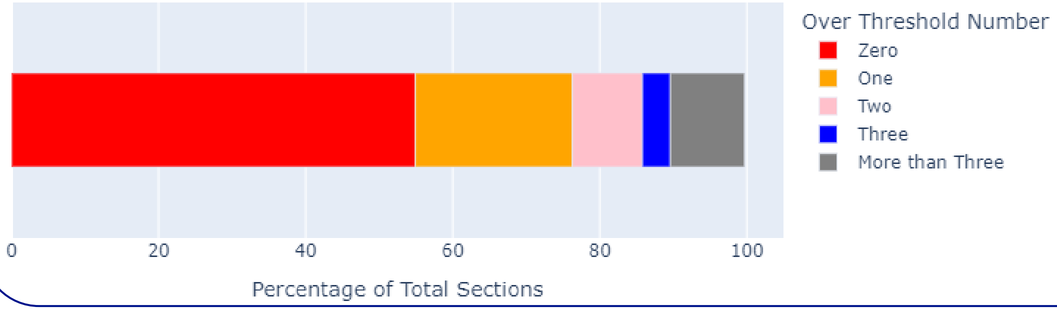
- Test Points for the Cathodic Protection Program **are grouped into Sections**
- Sections are defined by their Type and location (each defined within a single City/Town)
- Different **sections have different lengths, diameters, etc.** due to a variety of factors such as service area and when the pipe was installed
- **More than 20,000 Sections** in the dataset (~5,000 Sections with Annual Test Points)
- **More than 90,000 Test Points** (~22,000 Annual Test Points)

- **Corrosion Data**
  - Summary and basics of the structure
  - **First look at our critical variable**
  - Other key variables
- **Model Graveyard**
  - Original Plans
  - Regression & Survival Analysis
  - Pivot

The majority of test points have never failed an inspection



- Over 60% of **Individual Test Points** have never failed in ~22 years of data

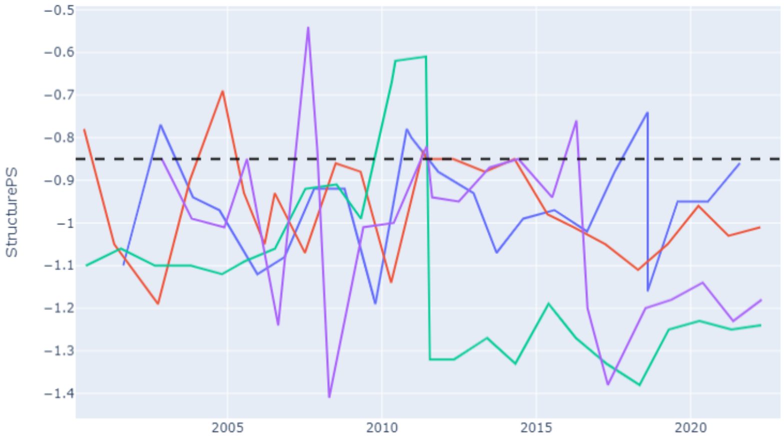


- Even at the **Section Level**, the majority have not had a failure in the observation period

Test point readings can exhibit stochastic behavior

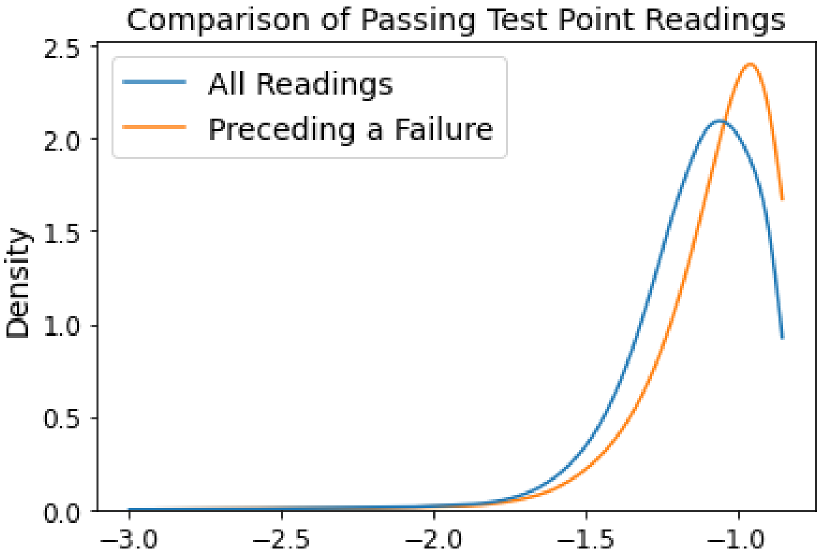
Test Points with No Failures

Test Points with At Least One Failure



- 4 Test Points chosen at random from each category to illustrate patterns
- Pipe to Soil Reading values and variance are both noisy
- Maintenance can play a role, but does not always

Readings prior to failure are hard to distinguish from normal

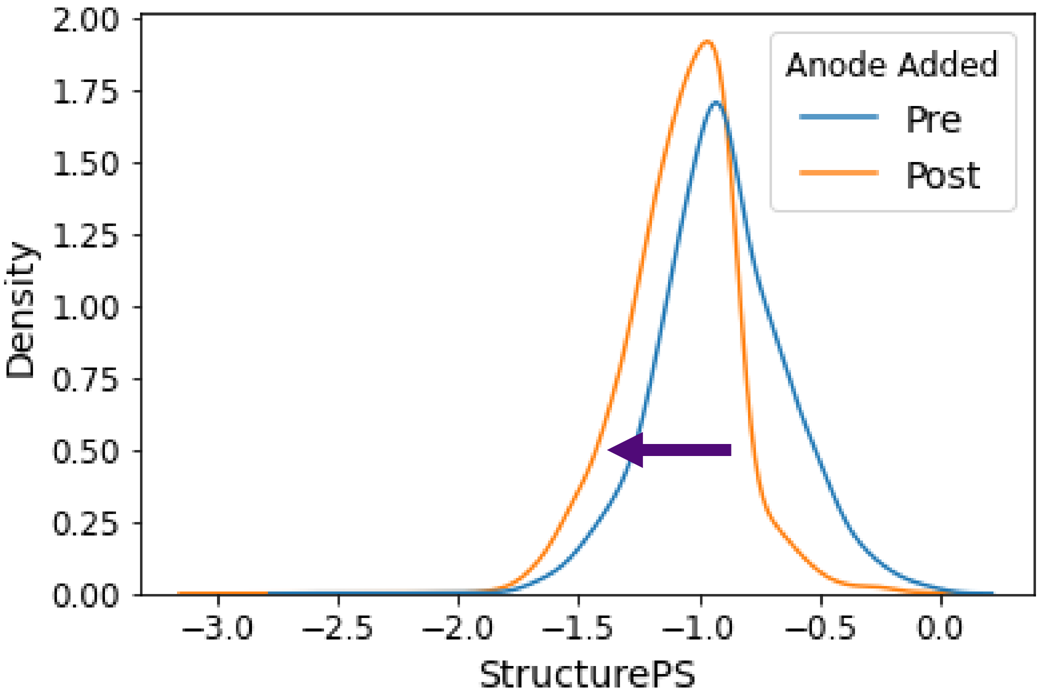


- The distribution of Test Point readings immediately preceding a failure is more concentrated around the threshold of -0.85
- There is substantial overlap, however, with the distribution of all Passing Readings

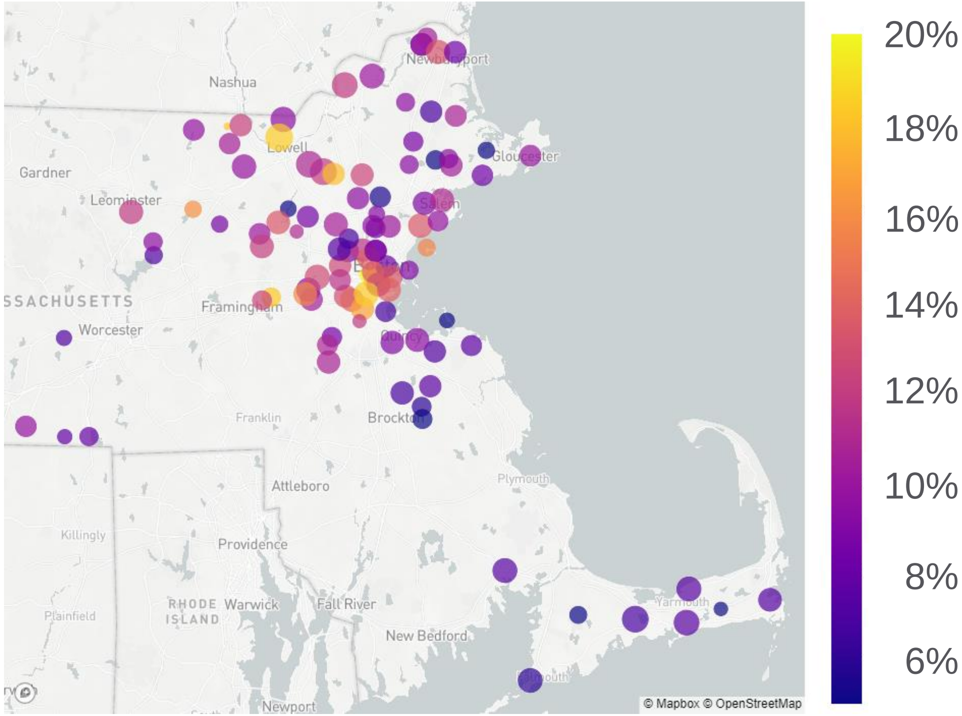
- **Corrosion Data**
  - Summary and basics of the structure
  - First look at our critical variable
  - **Other key variables**
- **Model Graveyard**
  - Original Plans
  - Regression & Survival Analysis
  - Pivot

Adding anodes moves readings away from threshold

- For Test Points where Anodes were added via Maintenance, readings were more negative after the work was complete
- Different patterns emerge for each of the many maintenance orders performed by teams in the field



Location Matters: Failure Percentage of All Readings by Town



- Test Points with the highest rates of failure are close to urban and industrial centers

- Possible factors per SMEs:
  - Age of Pipe
  - Pipeline Diameter
  - Proximity to the T*
  - Soil Moisture
  - …

*Boston's rapid transit system

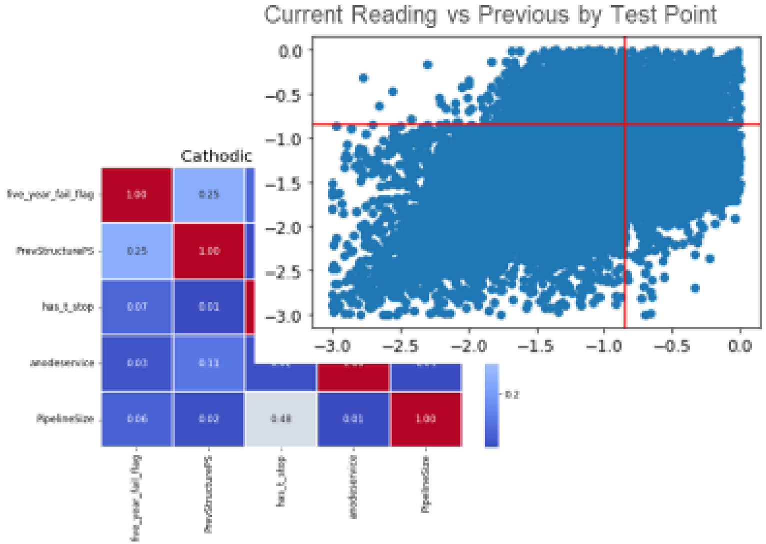
- **Corrosion Data**
  - Summary and basics of the structure
  - First look at our critical variable
  - Other key variables
- **Model Graveyard**
  - **Original Plans**
  - Regression & Survival Analysis
  - Pivot

Original project goals

- To support a wide-variety of initiatives and provide the most flexible outputs, the original project plan was to produce continuous outputs either by:
  - Predicting specific reading outputs over time
  - Building expected-time-to-fail metrics for each test point

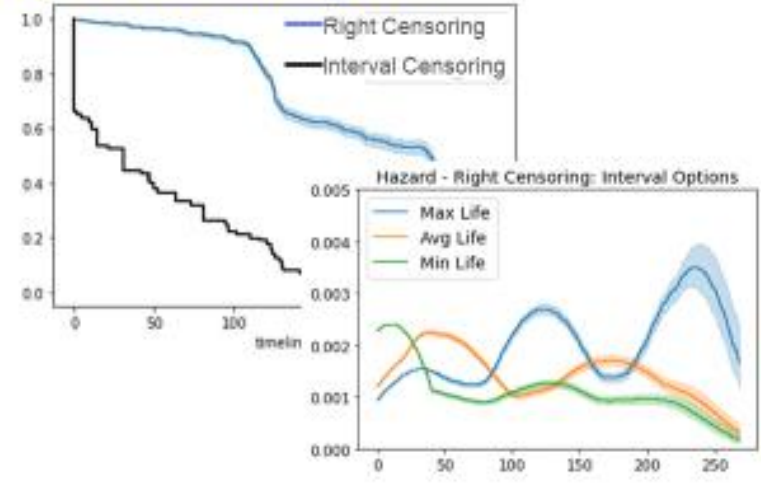
- **Corrosion Data**
  - Summary and basics of the structure
  - First look at our critical variable
  - Other key variables
- **Model Graveyard**
  - Original Plans
  - **Regression & Survival Analysis**
  - Pivot

Regression



- Limited history by test point, previous reading has limited predictive power
- Few variables exhibit any correlation with failed readings
- Impossible to support precision for estimates near -0.85

Survival Analysis



- Many test points unobserved for years, interval censoring approach misses out on 81% of data as they are Right Censored
- Hazard functions exhibit multi-modal behavior that suggest other mechanisms at play in terms of time-to-failure not captured by model

- **Corrosion Data**
  - Summary and basics of the structure
  - First look at our critical variable
  - Other key variables
- **Model Graveyard**
  - Original Plans
  - Regression & Survival Analysis
  - **Pivot**

Choosing a Modeling Path

Several modeling approaches were considered and eventually deemed unfit:

Efforts to use regression techniques to predict specific Pipe to Soil readings in the future failed to meet any standard of accuracy.

Survival Analysis, particularly a Proportional Hazards approach, helped us identify some important variables. In the end, however, the outputs were not answering the specific questions our partners needed.

In the end, a classification approach presented the best path forward. Our goal was to produce a 5-year probability of failure for every test point in the Massachusetts network

04

Model v1.0

Generating Probabilities

nationalgrid

Preparing the data

Classification target is whether a test point has *at least 1* failure within 5 years

- Take each test point reading as an **event**, capture relevant features
- Look 5 years ahead and see if that test point fails (Target, 1/0)
- This limits our dataset to readings taken before 2017
- Periodically test points are retired, so **events** without at least 2 readings in the following 5 years are removed

Our data are imbalanced

- Fewer than 1 out of 5 **events** in our dataset see a failure in the 5-year horizon
- Rebalancing and under sampling was considered to avoid over predicting 0 (since a naïve model would achieve >80% accuracy)

Choosing the right numbers

Ensuring model adoption and usefulness requires optimizing for the metrics end users care the most about and providing them with outputs that are useful and intuitive.

Optimization

- Resources are limited but all test points require observation and maintenance
- Being late to a failure on a "guaranteed street" presents big risk
- f1 & Balanced Accuracy

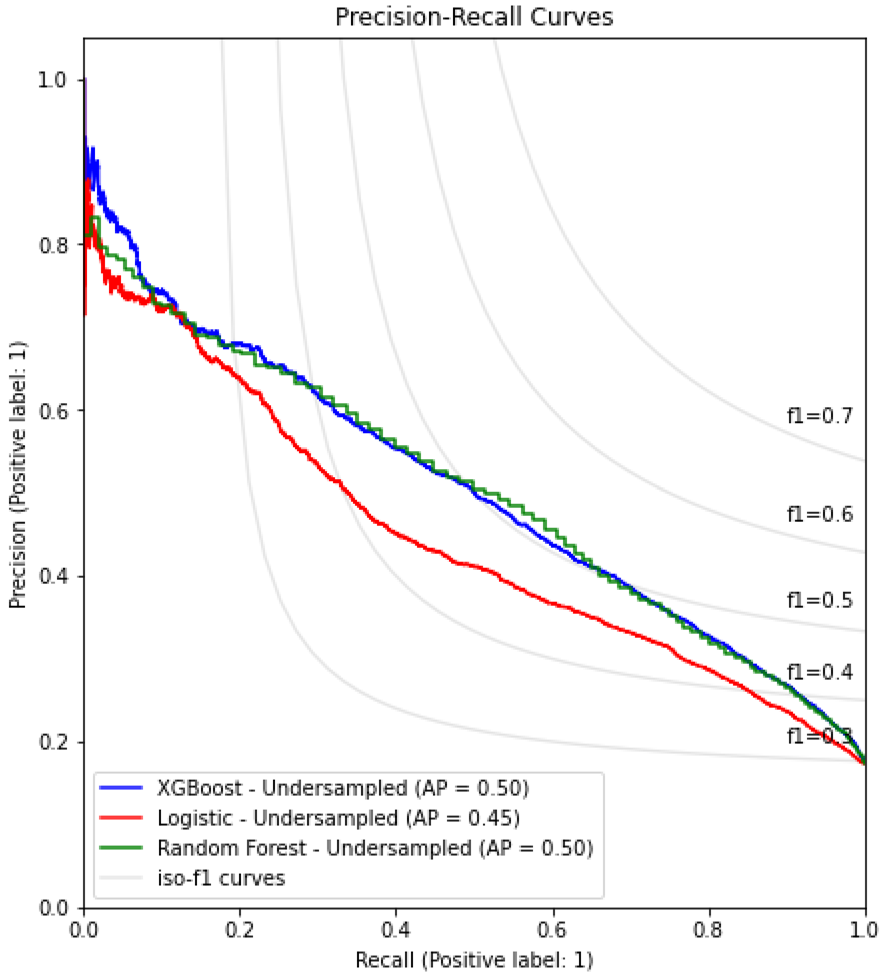
Outputs

- Ranking is helpful, but how does #10 compare to #20?
- Balancing proactive and reactive work orders along with other workstreams requires more context
- Probabilities of Failure

Model Comparison

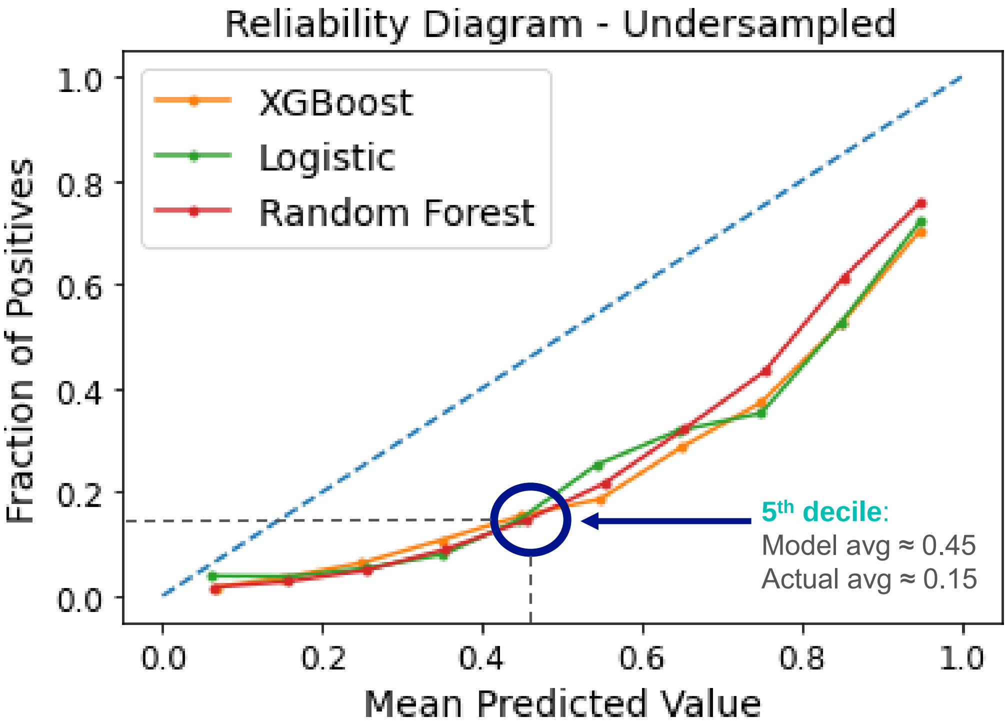
	Model	Accuracy	Balanced Accuracy	f1
	Dummy	0.8230	0.5000	0.0000
As-is	Logistic	0.8410	0.5940	0.3203
	Random Forest	0.8535	0.6369	0.4213
	XGBoost	0.8530	0.6452	0.4378
Rebalanced	Logistic	0.7401	0.6972	0.4616
	Random Forest	0.8513	0.6213	0.3868
	XGBoost	0.8530	0.6452	0.4378
Under Sampled	Logistic	0.7402	0.6973	0.4617
	Random Forest	0.7477	0.7368	0.5037
	XGBoost	0.7406	0.7332	0.4962

*Results above are averages from 5-fold Cross Validation

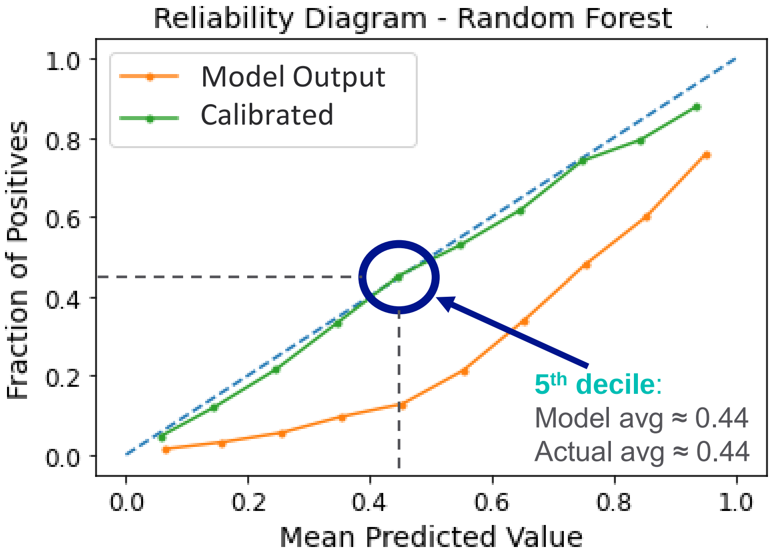


Classification scores to probabilities

- The outputs from a classification algorithm look a lot like probabilities, but using them as such will skew your distribution
- This chart shows increasing deciles of under-sampled **classification outcomes** and **compares** them **to observed failures**
- All 3 approaches drastically **overstate the likelihood of failure**

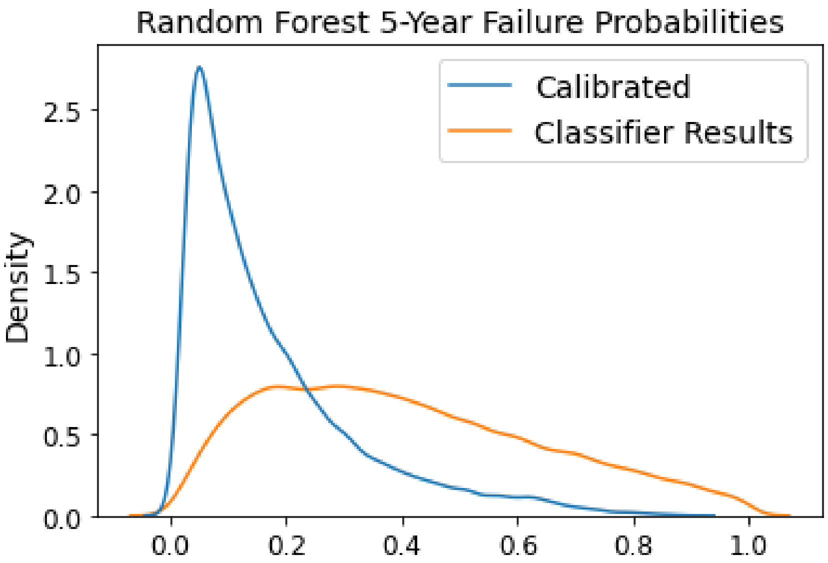


Classification scores to probabilities (cont.)



- To avoid over-stating probabilities, we calibrate the classification scores using Isotonic Regression
- Weighted least squares regression to transform *(Isotonic merely refers to increasing mapping of scores to probabilities)*
- The deciles in green show the calibrated values and actual results are far closer to unity (in blue)

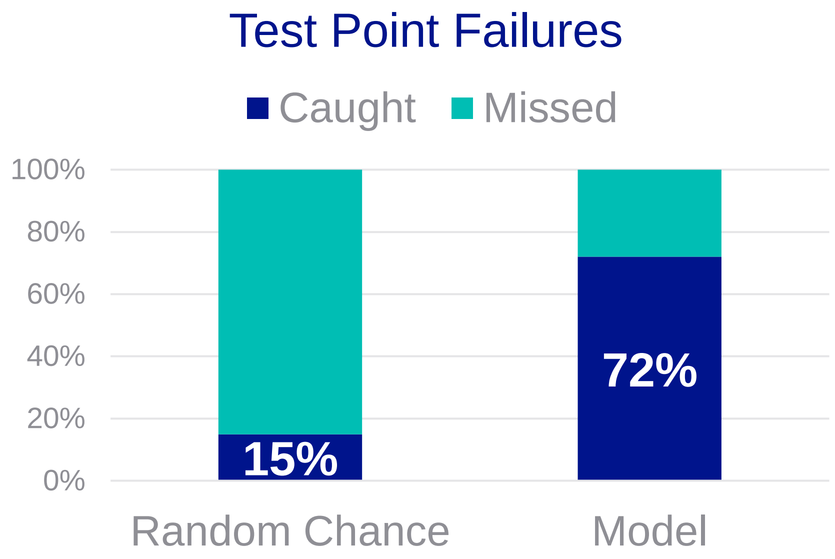
Classification scores to probabilities (cont.)



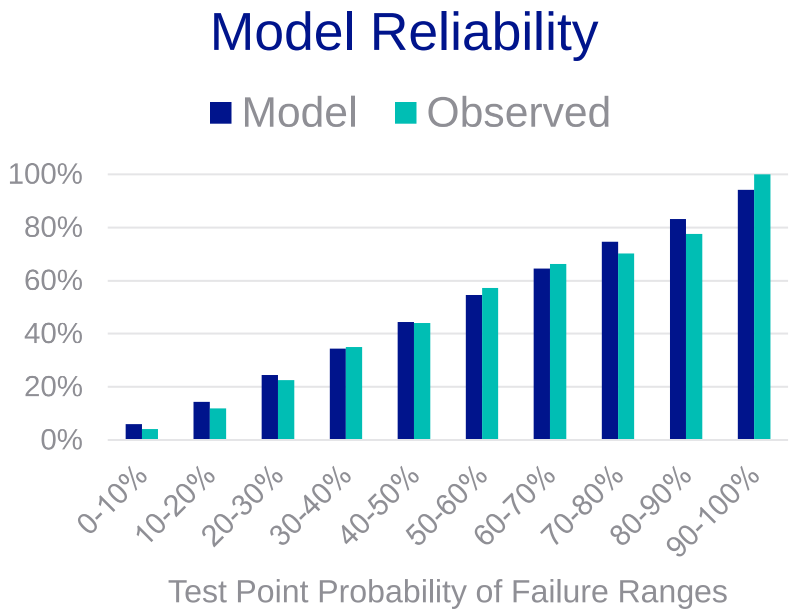
Dataset	5-Year Failure Percentage
Validation	18.0%
Training	17.8%
Test	17.4%
Classifier	41.0%
Calibrated Classifier Output	17.3%

- Without calibration, the failure rate for our test point population would be drastically overstated
- Calibrating the probabilities brings the cumulative forecast much closer to reality and **provides more useful direction to those planning work** and inspections

Performance on the Validation Set



Model correctly predicts 72% of failures

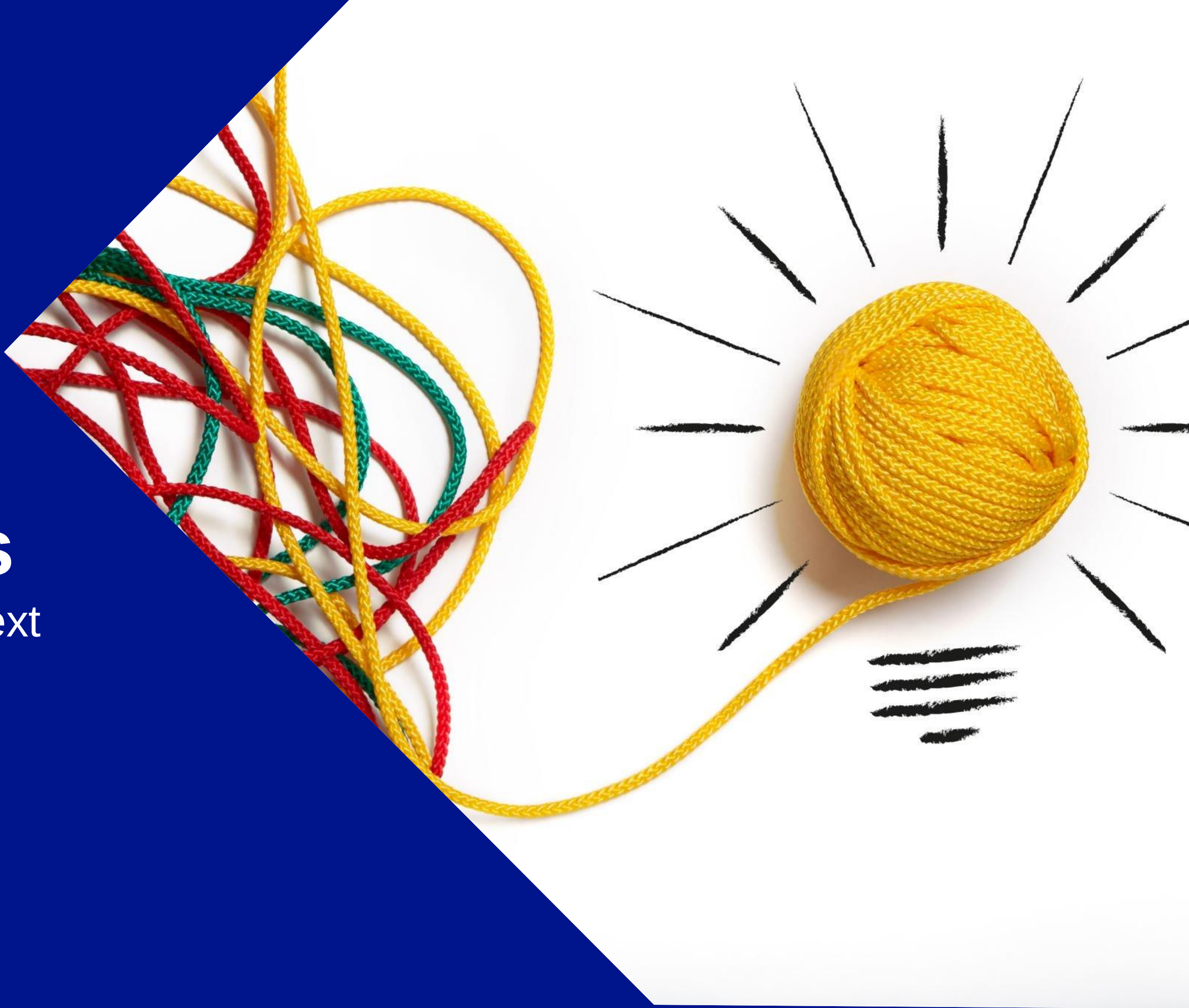


Model estimates match observations: Test Points modelled as failing 70% of the time *do* fail 70% of the time

05

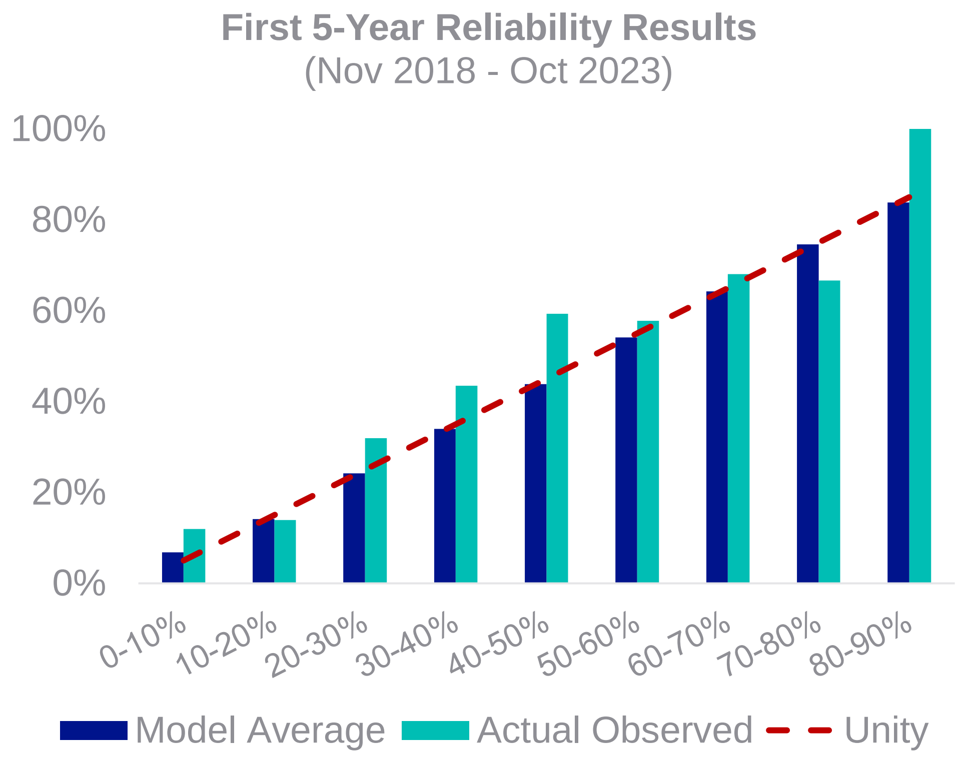
Current Status

Evaluating year one and next steps for the project



nationalgrid

Evaluating model performance after the first year



Full Period

- 18% of Test Points failed within the period overall, **for the top 100 risk-ranked test points 71% failed**

Calendar Year since kick-off

- Nearly 10% of failures in the first year of the program were in our top 1% of predicted Test Points
- Failed test points in Year 1 had a 50% higher average probability of failure than typical Test Points

Year one successes and next steps

Current Status

- Our partners in this project have begun utilizing the output to enhance their efforts to maintain compliance and prioritize work to support the Massachusetts gas service.
- “Proactive Maintenance” work orders submitted for the first time
- Improved relationships with cities and towns: Guarantee **backlog** (where we had been blocked from digging to address an issue) **down 50% in the first 6 months of use**

Next Steps

- Model evaluation over time
- Version 2 to explore other external datasets
- Recently kicked off similar effort in Upstate New York

nationalgrid