Risk-Based Decision-Making Supported by Machine Learning

by Michael P. Gloven, PE
Managing Partner, EIS

Pipeline Pigging and Integrity Management Conference

Marriott Marquis Hotel, Houston, USA
February 18-22, 2019

CLARION TECHNICAL CONFERENCES

Great Southern Press

Organized by
Clarion Technical Conferences *and* **Great Southern Press**

Introduction

As the pipeline industry continues to move towards achieving zero pipeline incidents[1] and underlying integrity data becomes more available and accessible, machine learning is emerging as a valuable practice to support the determination and validation of risk beliefs, measurement of root cause data and optimization of mitigation decision-making.

This paper presents the fundamental elements of machine learning as applied to linear and networked pipeline assets. Machine learning is simply a process to reveal useful patterns in data thru common methods found in linear algebra, descriptive and inferential statistics, and calculus. Underlying threat susceptibility and severity models are learned and validated through actual observations. The models then support the assessment of un-piggable pipelines, dig prioritization programs, inference of missing data, prioritization of data collection activities, analysis of interactive threats, optimization of inspection intervals and selection of mitigative actions. More importantly, data-driven learned models are explicitly validated through actual observations, an often-overlooked concept in existing risk practices.

Case Study

We're interested in learning models based on known external corrosion observations and then applying these models to assets of similar types to predict the potential susceptibility and severity of external corrosion.

The case study demonstrates how two machine learning categories, classification and regression, work together to predict external corrosion susceptibility and severity, respectively. Both categories are managed thru the same machine learning process yet have important differences in their observational inputs, learning methods, performance vectors and outputs.

Susceptibility models predict a level of confidence or probability of the expected presence or non-presence of external corrosion thru the use of supervised classification methods. Severity models predict the potential depth or rate of external corrosion thru the use of supervised regression methods.

The results of the models are combined to answer the question of threat susceptibility or exposure (i.e. predicting susceptibility as 0-100% confidence or probability) and, current and future severity (i.e. predicting wall loss depth over time). The models are applied to dynamically segmented pipes of similar types to predict expected external corrosion events (failures) over time. The remainder of this paper demonstrates how this is accomplished.

Machine Learning Process

Figure 1 summarizes a supervised machine learning process. Learning or observational data trains one or more methods which become models validated thru "held-back" observations. Models meeting domain expert criteria and performance are deployed to make predictions on assets of similar type.

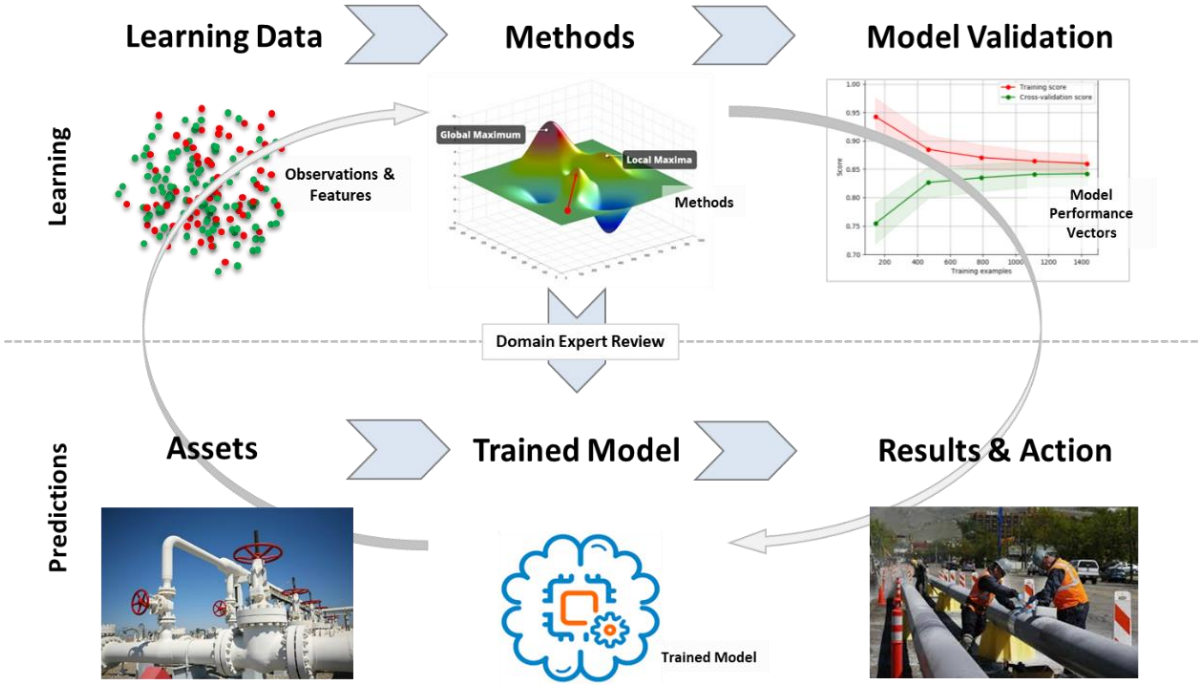


Figure 1 - Machine Learning Process

The overall process is used to support important safety and reliability objectives:

- Identify or Predict Threat Susceptibility & Severity
- Analyse Interactive Threats
- Measure the Value of Root Cause Data
- Plan & Optimize Inspections & Assessments
- Validate Risk Algorithms & Beliefs
- Support & Validate Mitigative Decision-Making

1. Define Objective

The first step in the machine learning process is to identify a clear objective. In this case we're interested in learning models to apply to similar yet unseen dynamically segmented pipes to predict corrosion susceptibility (0-100% probability) and severity (0-100% wall loss depth over time). We want to select models which have an acceptable level of performance and are capable of supporting risk mitigation decision-making.

2. Define & Collect Observations

Collecting learning data requires expert understanding of what defines an observation and the concept of optimal sample size. Domain experts normally define what constitutes a valid observation by considering measurement thresholds and relevant windows of time. Measurement thresholds are often constrained by the measurement technology whereas the validity of aged data is determined by domain experts. Optimal sampling may be achieved thru a validated random sampling process, consideration of the central limit theorem and learning curves.

Classification Observations

Figure 2 is an example of classification observations considering relevant time, dynseg lengths and detection tolerance. Classification model learning requires binominal (true\false) or polynominal observations of the targets of interest across the learning data set. Classification observations are used to learn susceptibility models.

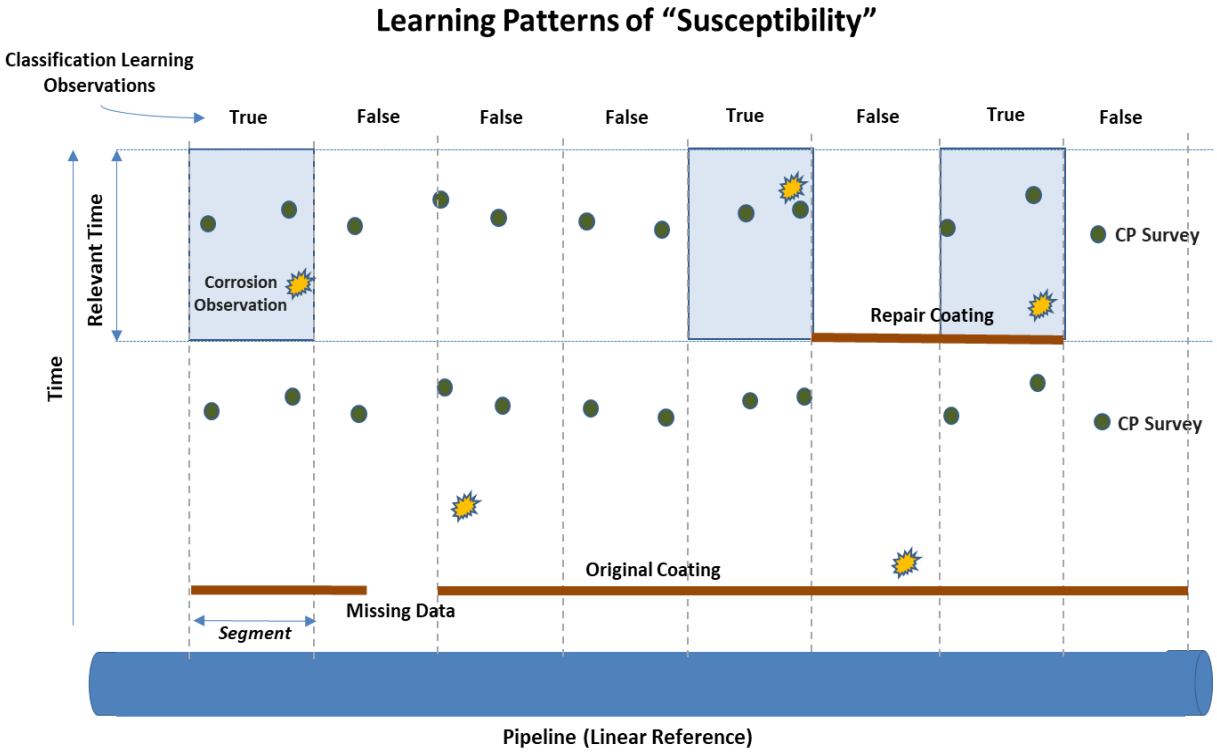


Figure 2 - Classification Learning Data

Regression Observations

Figure 3 is an example of regression observations. In this case, only dynseg segments or points with observed wall loss measurements are considered as observations for learning. Regression observations are used to learn severity models.

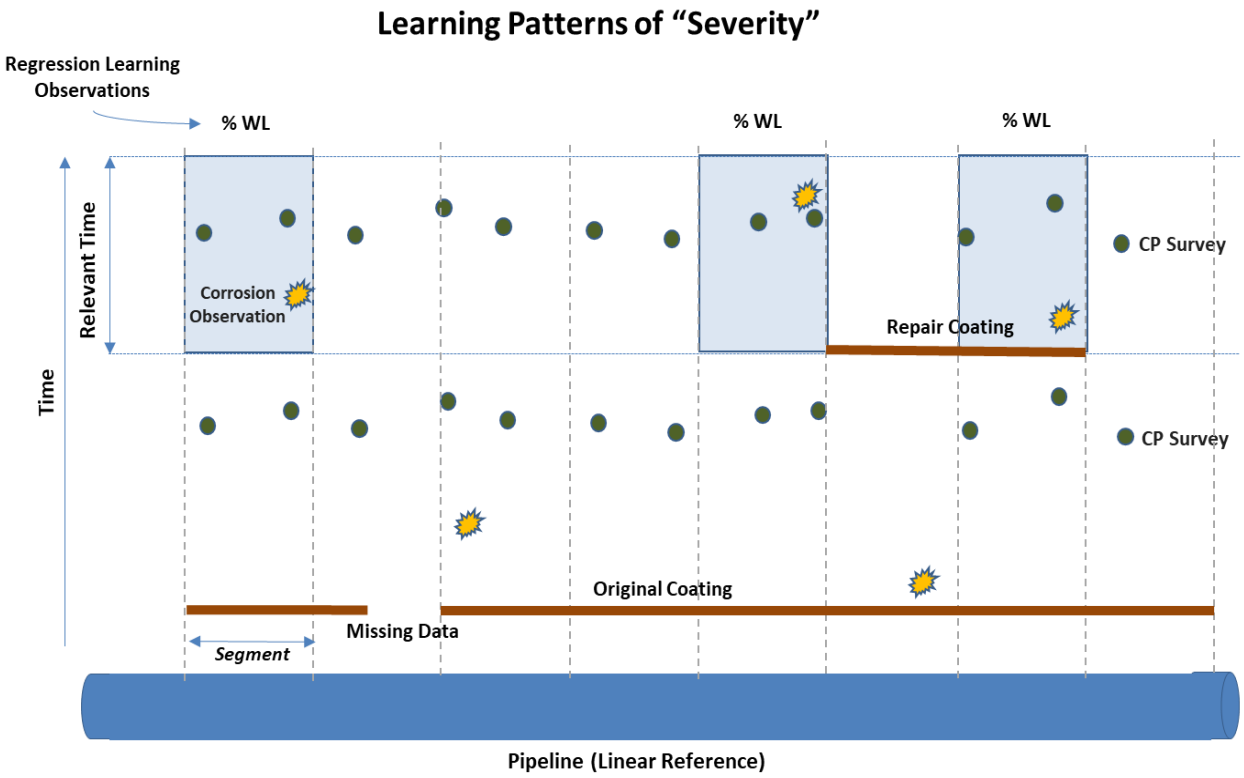
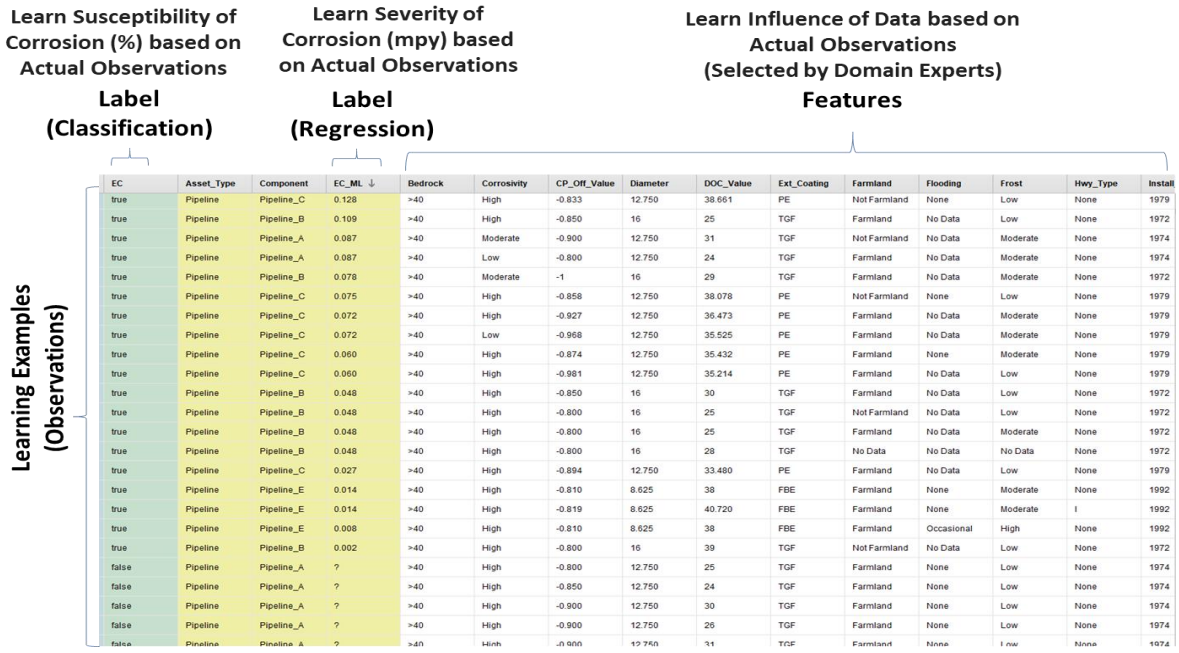


Figure 3 - Regression Learning Data

Aggregate Observational Data

As shown in figure 4, the practitioner combines the classification and regression observations to create a learning data set of records (vector or tensor) with true\false indications and numerical



EC	Asset_Type	Component	EC_ML	Bedrock	Corrosivity	CP_Off_Value	Diameter	DOC_Value	Ext_Coating	Farmland	Flooding	Frost	Hwy_Type	Install
true	Pipeline	Pipeline_C	0.128	>40	High	-0.833	12.750	38.661	PE	Not Farmland	None	Low	None	1979
true	Pipeline	Pipeline_B	0.109	>40	High	-0.850	16	25	TGF	Farmland	No Data	Low	None	1972
true	Pipeline	Pipeline_A	0.087	>40	Moderate	-0.900	12.750	31	TGF	Not Farmland	No Data	Moderate	None	1974
true	Pipeline	Pipeline_A	0.087	>40	Low	-0.800	12.750	24	TGF	Farmland	No Data	Moderate	None	1974
true	Pipeline	Pipeline_B	0.078	>40	Moderate	-1	16	29	TGF	Farmland	No Data	Moderate	None	1972
true	Pipeline	Pipeline_C	0.075	>40	High	-0.858	12.750	38.078	PE	Not Farmland	None	Low	None	1979
true	Pipeline	Pipeline_C	0.072	>40	High	-0.927	12.750	36.473	PE	Farmland	No Data	Moderate	None	1979
true	Pipeline	Pipeline_C	0.072	>40	Low	-0.968	12.750	35.525	PE	Farmland	No Data	Moderate	None	1979
true	Pipeline	Pipeline_C	0.060	>40	High	-0.874	12.750	35.432	PE	Farmland	None	Moderate	None	1979
true	Pipeline	Pipeline_C	0.060	>40	High	-0.981	12.750	35.214	PE	Farmland	No Data	Low	None	1979
true	Pipeline	Pipeline_B	0.048	>40	High	-0.850	16	30	TGF	Farmland	No Data	Low	None	1972
true	Pipeline	Pipeline_B	0.048	>40	High	-0.800	16	25	TGF	Not Farmland	No Data	Low	None	1972
true	Pipeline	Pipeline_B	0.048	>40	High	-0.800	16	25	TGF	Farmland	No Data	Moderate	None	1972
true	Pipeline	Pipeline_B	0.048	>40	High	-0.800	16	28	TGF	No Data	No Data	No Data	None	1972
true	Pipeline	Pipeline_C	0.027	>40	High	-0.894	12.750	33.480	PE	Farmland	No Data	Low	None	1979
true	Pipeline	Pipeline_E	0.014	>40	High	-0.810	8.625	38	FBE	Farmland	None	Moderate	None	1992
true	Pipeline	Pipeline_E	0.014	>40	High	-0.819	8.625	40.720	FBE	Farmland	None	Moderate	I	1992
true	Pipeline	Pipeline_E	0.008	>40	High	-0.810	8.625	38	FBE	Farmland	Occasional	High	None	1992
true	Pipeline	Pipeline_B	0.002	>40	High	-0.800	16	39	TGF	Not Farmland	No Data	Low	None	1972
false	Pipeline	Pipeline_A	?	>40	High	-0.800	12.750	25	TGF	Farmland	None	Low	None	1974
false	Pipeline	Pipeline_A	?	>40	High	-0.850	12.750	24	TGF	Farmland	None	Low	None	1974
false	Pipeline	Pipeline_A	?	>40	High	-0.900	12.750	30	TGF	Farmland	None	Low	None	1974
false	Pipeline	Pipeline_A	?	>40	High	-0.900	12.750	26	TGF	Farmland	None	Low	None	1974
false	Pipeline	Pipeline_A	?	>40	High	-0.900	12.750	31	TGF	Farmland	None	Low	None	1974

Figure 4 - Complete Learning Data Set

values of wall loss. These observations are called "labels" indicated as "EC" (external corrosion) and EC_ML (external corrosion metal loss). Underlying potential root cause and correlation data identified by domain experts are shown as columns of "features". The resulting matrix is used to learn models.

Sample Optimization

The optimal or minimal number of observations required depends on the objective and may be determined thru inferential statistical methods or learning curves. Learning curves support an iterative process of model development by plotting model validation and testing performance versus sample size. The point on the learning curve where model performance or error is minimized may be considered the point of best sample size.

As shown in figure 5, for this case study a sample size of 8000 observations is best considering R2 regression performance (coefficient of determination). The practitioner has several performance vectors to choose from depending on the model method and overall objective. Learning curves provide an iterative structure to converge on best sample sizes for the given problem.

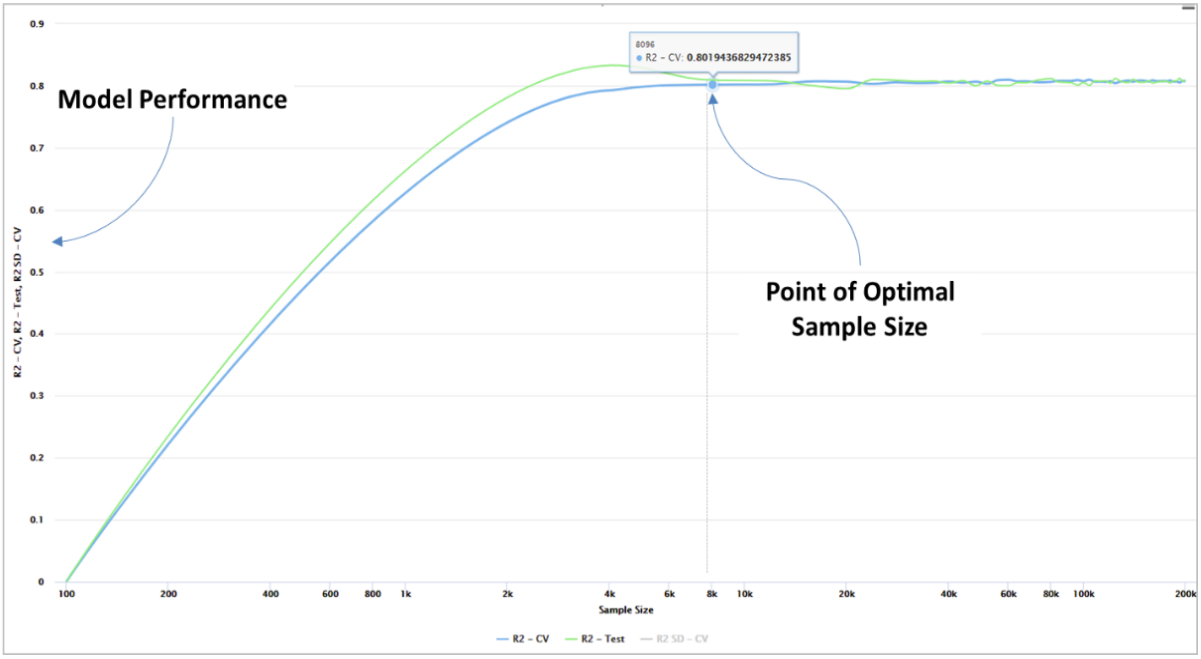


Figure 5 - Determining Sample Size thru Learning Curves

3. Define & Collect Influencing Attributes

Influencing causation data or "features" applicable to the target of interest are typically defined by domain experts. Once collected, quality assurance processes are applied to assess accuracy, missing data, temporal measurements, locational accuracy, alignment of sample data with the population, potential outliers, data stability and collinearity. A reliable and complete data set corrects relevant quality issues.

Measurement of Data Importance

Now that a data set is complete, the practitioner may be interested in knowing how the underlying data influences external corrosion prior to formal machine learning. Having an understanding of data

influence may drive data prioritization or additional data collection activities, or removal of unnecessary data from the analysis.

Numerous methods are available to inform the practitioner of data priority and impact on the target of interest. Correlation, information gain, deviation and evolutionary methods are common approaches.

Each method is simply a statistical or mathematical approach to measure the relationship of underlying data to the target of interest. As shown in figure 6, results are normalized, summed and sorted by feature. Introducing a random variable into the analysis may indicate which data is no better than a coin-flip as far as influence on the target.

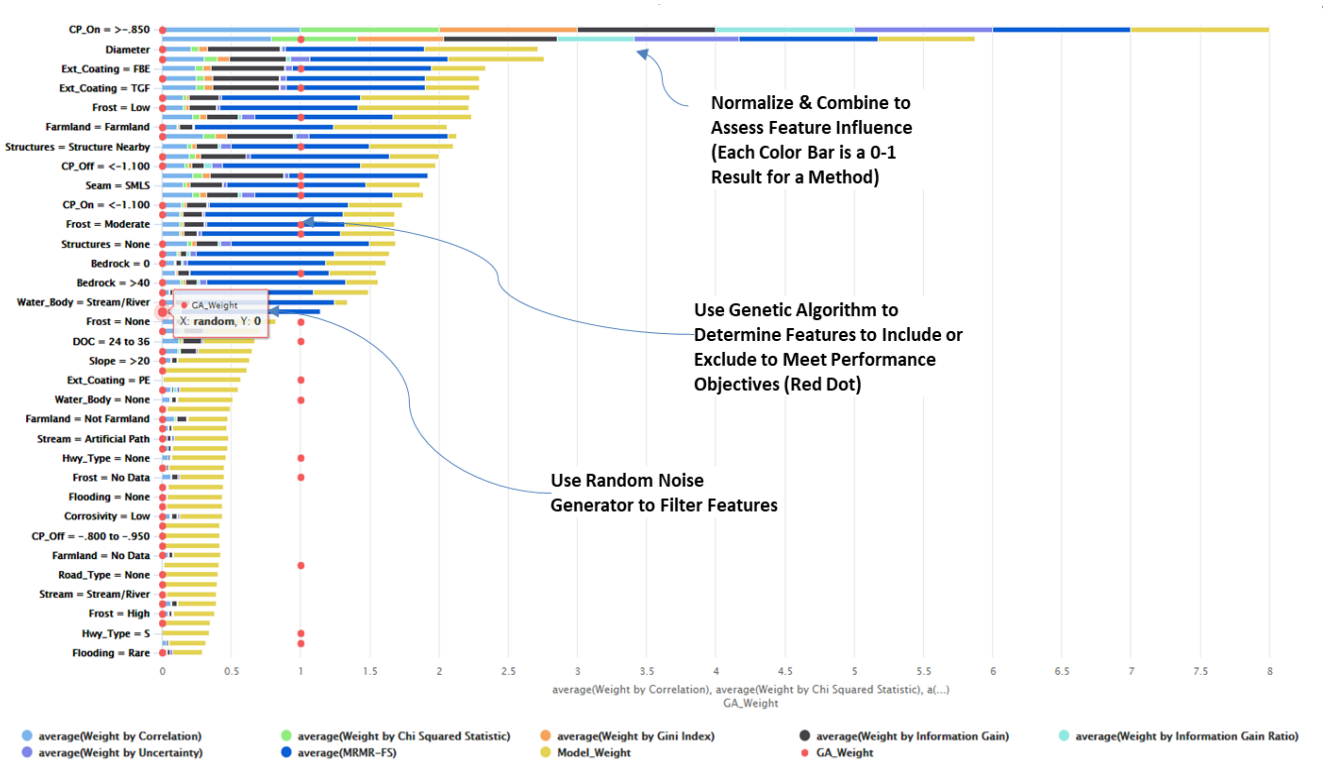


Figure 6 - Feature Influence on External Corrosion

Note that CP On, Diameter and Coating Types track positively with the presence and severity of external corrosion, whereas the presence of Highway crossings has limited impact and its information is no better than a random coin flip. These insights may be used to guide the practitioner on data prioritization and investment.

Measuring the contribution of attribute data based on model performance may also inform the practitioner if additional data is required or if there is too much data. As with sampling, this involves the use of learning curves and iterative model validation, and a combination of evolutionary feature selection.

Figure 7 shows the results of evolutionary analysis and the minimal\optimal data set to achieve an acceptable model performance for external corrosion. In this example, the best model accuracy is achieved by considering eleven features. The example also demonstrates the impact on performance when reducing the number of features.

Features	Names	accuracy
	Corrosivity, CP_Off, CP_On, Diameter, Farmland, Flooding, Frost, Power_Line, Seam, Slope, Water_Body	0.823
11	Corrosivity, CP_Off, CP_On, Diameter, Farmland, Flooding, Frost, Power_Line, Seam, Slope, Water_Body	0.823
9	Corrosivity, CP_Off, CP_On, Diameter, Farmland, Flooding, Frost, Power_Line, Slope	0.822
8	Corrosivity, CP_Off, CP_On, Diameter, Farmland, Flooding, Frost, Power_Line	0.821
7	CP_Off, CP_On, Diameter, Farmland, Frost, Power_Line, Slope	0.807
7	CP_Off, CP_On, Diameter, Farmland, Frost, Power_Line, Slope	0.807
6	CP_Off, CP_On, Diameter, Frost, Power_Line, Slope	0.781
6	CP_Off, CP_On, Diameter, Frost, Power_Line, Slope	0.781
5	CP_Off, CP_On, Diameter, Farmland, Frost	0.775

Figure 7 - Evolutionary Feature Selection

Feature Engineering

Feature engineering is a common approach to aggregate large numbers of features into fewer dimensions. This is inherent with most sample data as features are rarely first principal based but rather aggregated in terms of cp level, soil type, coating, etc. The practitioner has the option of further aggregating these features into principal components or new dimensions to improve model accuracy, interpretation or run-time performance.

4. Learn & Optimize Methods

Creating the learning data set is an iterative process with the objective of achieving specific quality metrics, optimal sample size, and a set number of underlying feature or engineered feature attributes to meet model performance requirements.

Once the data set is ready, the practitioner may select numerous types of classification and regression methods[2] (figure 8) to learn susceptibility and severity models. Each method may perform differently for the given objective. A key role of the practitioner is to select the best method for the given situation.

Figure 8 - Learning Methods

Cross-validation and testing processes may be applied to each method using the learning data set. Cross validation is a method of learning a model with, let's say, 70% of the data. Test validation is then applying the 30% held-back data to the model. Both cross validation and testing deliver performance vectors in the form of a confusion matrix for classification and RMSE (root mean squared error) and R2 (coefficient of determination) for regression. Other performance vectors are available to the practitioner.

The case study initially applied eight classification methods to the learning data set. Figure 9 indicates learned model accuracies ranged between 75.7% and 82.8% with Deep Learning, Random Forest and Gradient Boosted Trees performing best. Sensitivities (minimizing model misses) ranged between 0% and 75.6% with Generalized Linear Model and Deep Learning performing best.

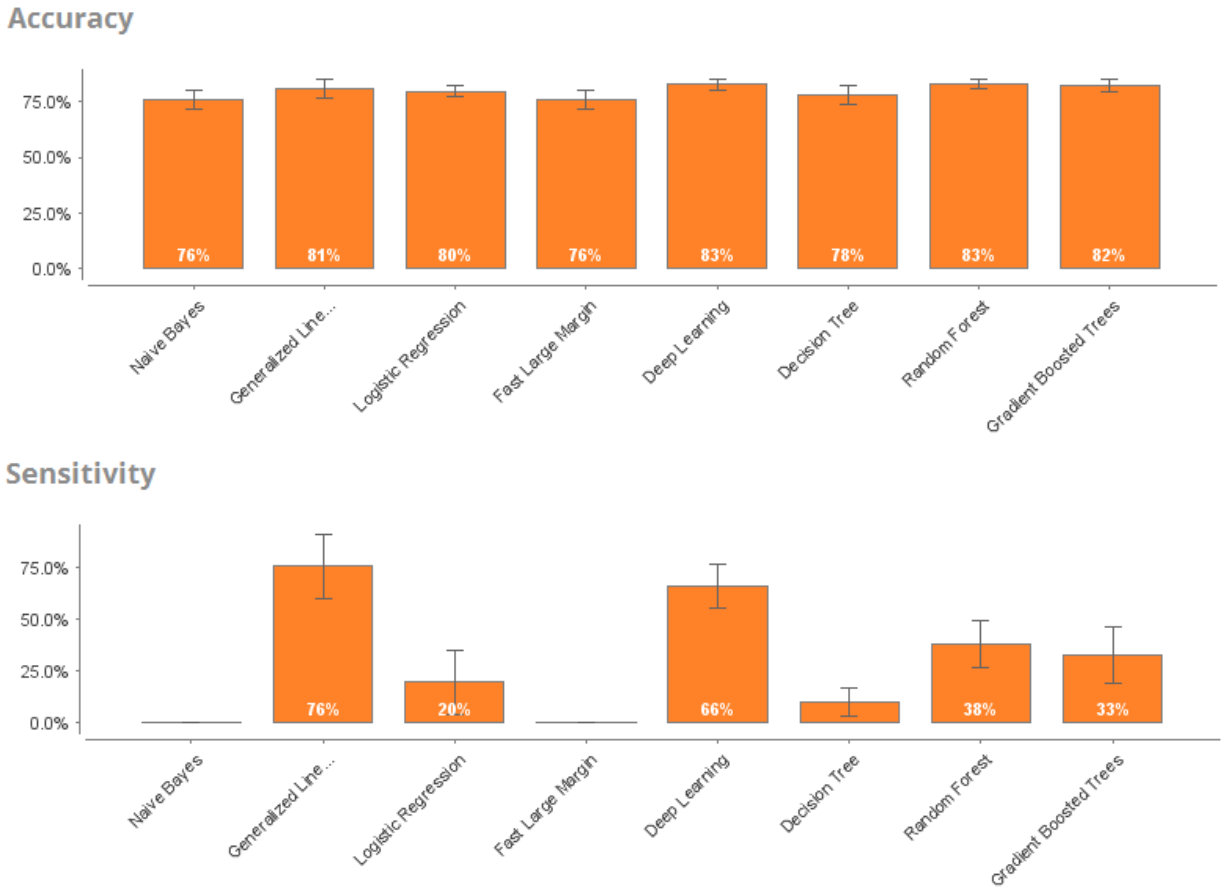


Figure 9 - Classification Performance

Classification model performance is often represented thru a confusion matrix as shown in figure 10. The confusion matrix indicates how the model predicts versus the actual observation. The simplified example within the confusion matrix demonstrates the potential trade-offs between incorrect prediction types for 100 segments of observed pipe, 10 with defects and 90 without. Selecting or tuning a model considers the strategic differences between accuracy, false positives and misses (sensitivity). Note the consistency in accuracy yet high variance of sensitivity with the case study models in figure 9.

Overall Accuracy 89%	Actual (No Defect in Pipe - 90)	Actual (Defect in Pipe - 10)	
Prediction (No Defect in Pipe - 81)	**80** (true negatives)	**1** (misses or false negatives)	
Prediction (Defect in Pipe - 19)	**10** (false positives)	**9** (true positives)	**47%** (precision)
Example - 100 Joints of Pipe	**89%** (specificity)	**90%** (sensitivity or recall)	

Figure 11 - Confusion Matrix

The case study initially applied five regression methods to the observation data set. Figure 11 indicates learned model RMSE between 7.5% and 8.7% wall loss and R Correlation between .125 and .490. The best R2 is .490^2 = .240 which means the underlying data explains 24% of the change in the prediction, hence 76% is unexplained.

Figure 10 - Regression Performance

The case study is seeking a model with +90% accuracy, a miss rate at least 100 times less than the false positive rate, R2 > .500 and an RMSE < 10% with a narrow standard deviation. Each of these criteria are considered in the final model selection. Determining the best model is an iterative process using learning curves and optimization techniques to determine optimal method hyper-parameters, sample representations and feature sets to meet the objective.

Model Error

At a fundamental level, the performance error of models is normally expressed as a combination of bias and variance. A role of the practitioner is to minimize and balance these errors[3]. As shown in figure 12, high bias error may be represented when both training and cross validation curves depart from the expected optimal (i.e. score expected to be closer to 1.00). High variance error is represented when the curves do not converge (not shown). An optimal scenario for a model is where the curves converge at expected optimal.

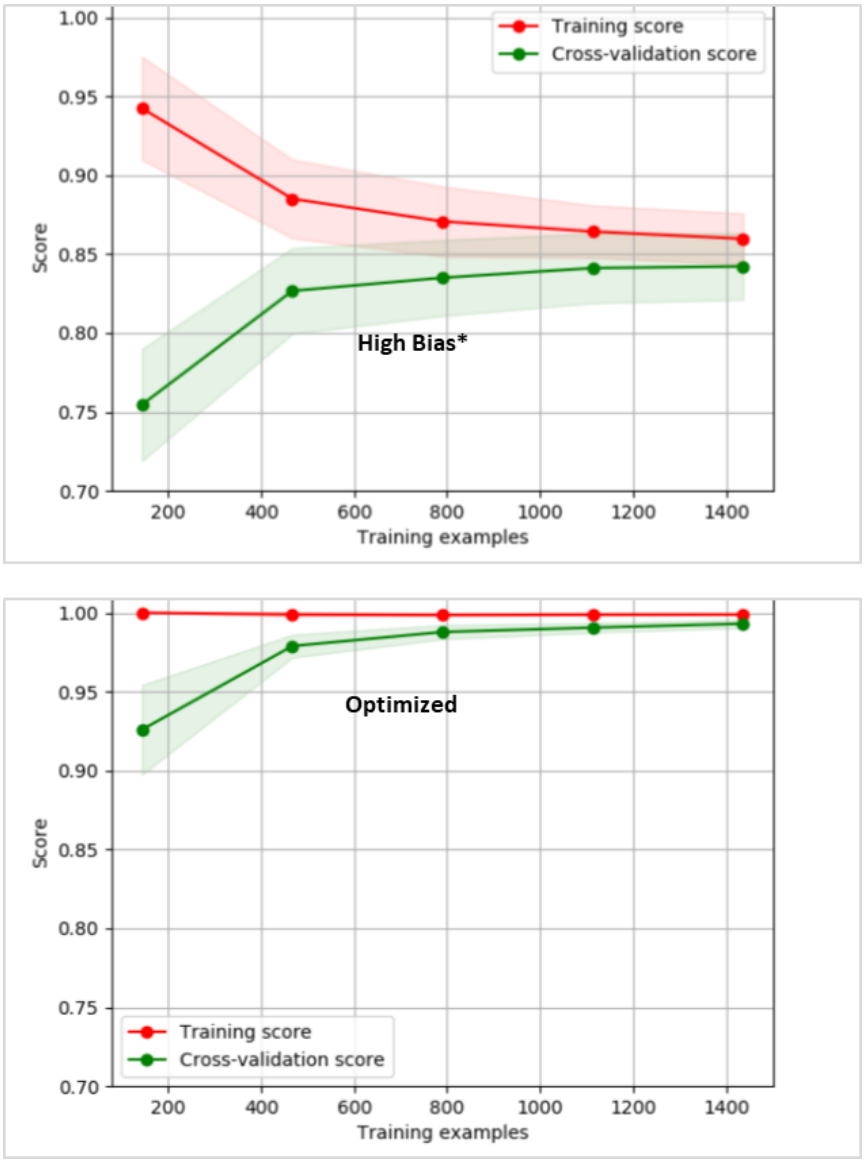


Figure 12 - Performance Learning Curves

High model bias may be corrected by introducing new attributes and additional sample data whereas high variance may be corrected by reducing the number of attributes and using model regularization techniques.

5. Select Best Models

Selecting the best model depends on the overall objective, how well the model performances, the cost of maintaining and running the model and how well it can be explained to stakeholders. After tuning hyper-parameters, adjusting sample representation and feature selection, the case study selected gradient boosted trees for both external corrosion susceptibility and severity.

Gradient boosted tree methods may be used for both classification and regression problems. The model is a series of decision trees which continue to improve learning on weak predictions based on practitioner criteria. Figures 13 and 14 show the final classification and regression models, respectively. Note the models are series of trees and the figures show just the first tree representation. The classification model predicts true\false external corrosion whereas the regression model predicts wall loss depth.

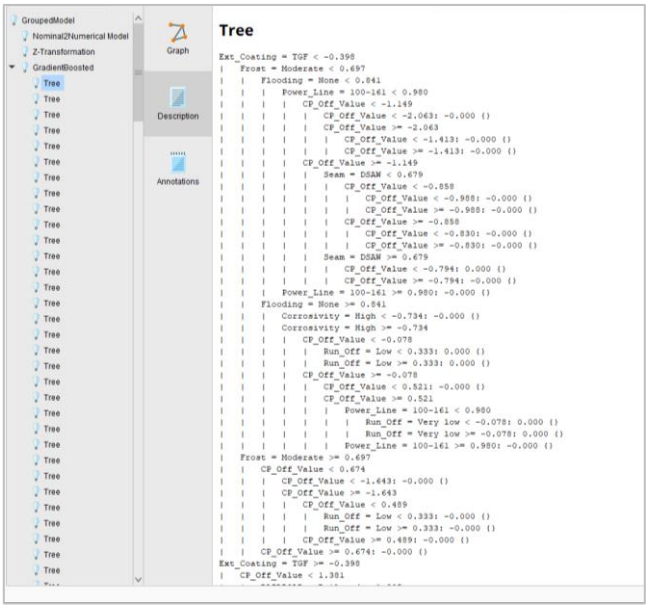


Figure 13 - GBT Classification

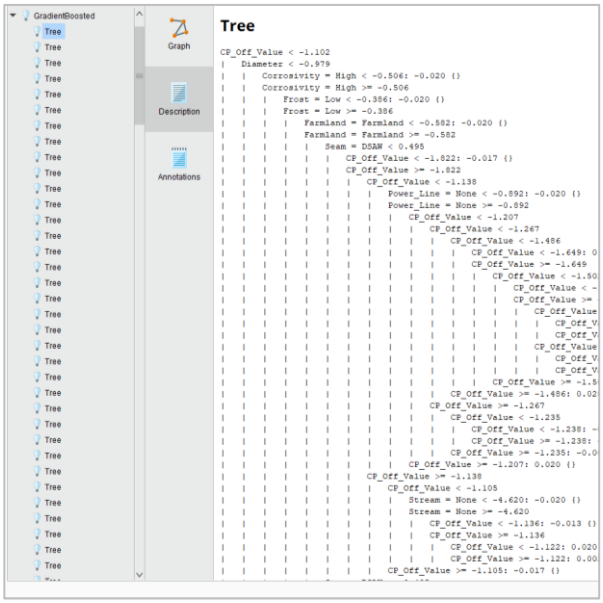


Figure 134 - GBT Regression

Figure 15 is the confusion matrix for the classification model. The regression performance vector RSME and R2 is +/- 9% and .547, respectively. The performance of both models meet project objectives.

accuracy: 94.17% +/- 0.48% (micro average: 94.17%)

	true false	true true	class precision
pred. false	16333	1424	91.98%
pred. true	543	15452	96.61%
class recall	96.78%	91.56%	

Figure 14 - Classification Confusion Matrix

6. Apply Model

Model application is performed thru a scoring process, that is, applying the learned models against dynamically segmented assets of similar types. Ensuring assets are of similar types involves verifying meta-data is common between learned and scored data, and the populations are similar.

One method of population similarity review is thru a t-SNE method where all data is generalized into 2 or 3 dimensions. We'd expect the learning and scoring populations to be contained in the same space. If the populations are in different spaces the practitioner should review the learning data to ensure it is representative of a random sample of the larger scored population. Figure 16 visualizes a two-dimensional t-SNE where the red learning data represents well into the green scoring data.

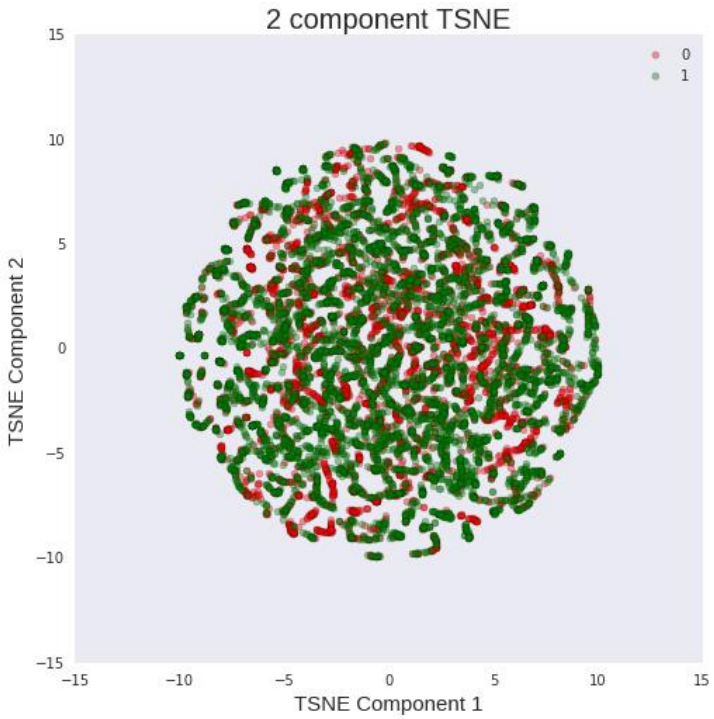


Figure 15 - t-SNE Chart

7. Combine Models & Results Analysis

The final step is applying the learned models to pipelines to predict external corrosion susceptibility and severity.

Scored classification (susceptibility) results are expressed as a probability or confidence of the presence of external corrosion on dynamically segmented pipe. To further understand the predictions, scored results for each pipe segment are considered as a probability of an event rate (PER), that is, a 0% PER indicates historical observations support a near zero chance of an external corrosion event whereas a 100% PER indicates a maximum number of events could occur. Learned PER's are based on industry and company specific incident rates and are thus grounded in historical observations.

The regression results are expressed as a severity or expected rate of degradation or mpy (expected mils per year). Regression results support the determination of "when" an event could occur and is calculated by determining TTF (time to failure) based on predicted wall loss, wall thickness, and age. The results are predicted as PoE's (probability of events) over time thru a two-parameter Weibull analysis[4].

Both classification PER and regression mpy, TTF and PoE results are combined to predict external corrosion events and support assessment and mitigative decision-making. PER is used to predict the current level of susceptibility and PoE is used to predict likelihood of events over time.

Since the models are data driven and held-back data is used to validate performance, each model and subsequent prediction results are associated with the model performance vectors. Thus, the results are characterized with specific levels of confidence, uncertainty and explainability. In addition, models are relearned and improved as new observations become available. The learning process is a continuously improving process.

Figure 17 illustrates external corrosion classification % PER values for a section of pipeline mapped with root cause data determined thru learning correlation methods. This example indicates root cause data (Structures, CP Off) influencing external corrosion susceptibility at station 82,960 ft. Thus, the results of the machine learning process indicate both level of susceptibility and potential root cause mitigation insights.

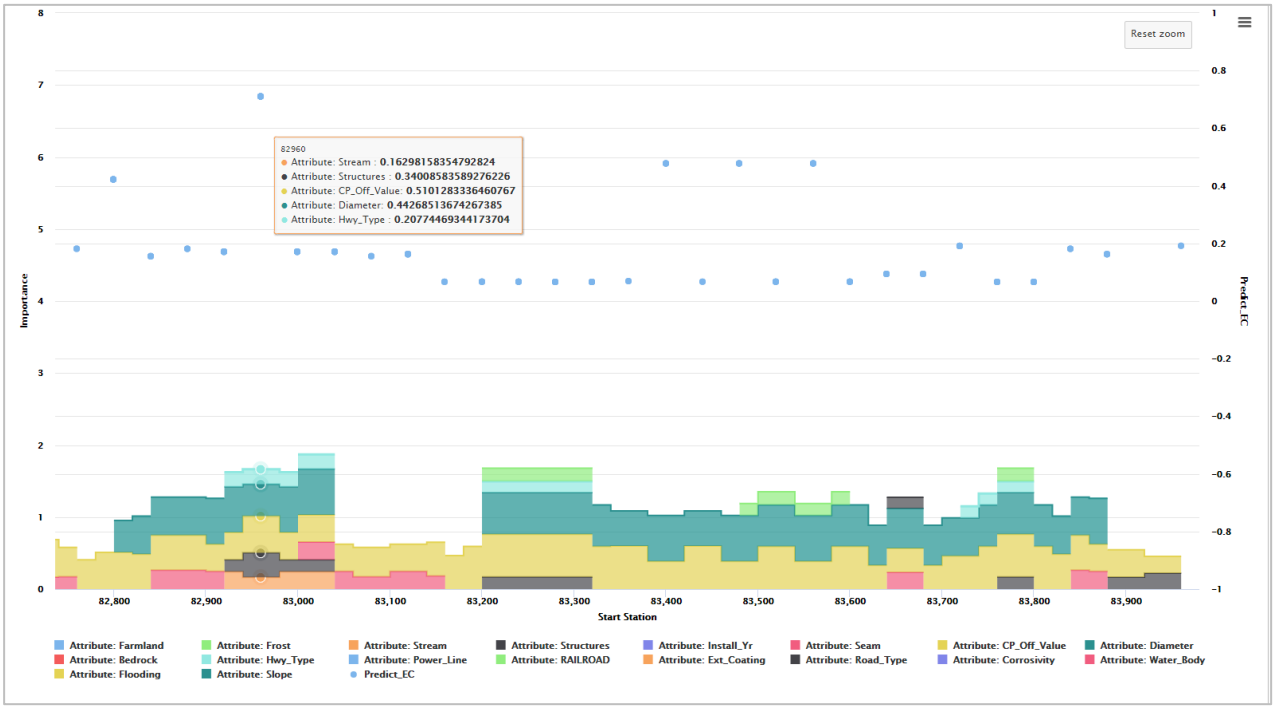


Figure 16 - % PER Results

Figure 18 illustrates external corrosion mpy regression analysis results for the same section of pipeline. These results are used to determine an estimated time to failure (TTF) then converted to a Probability of Event (PoE) thru a Weibull analysis. The combination of susceptibility based % PER and severity based PoE support the prediction of event rates over time as an event rate per segment of pipe for each year. These values may then be monetized thru integration of consequence data to support project investment decision-making.

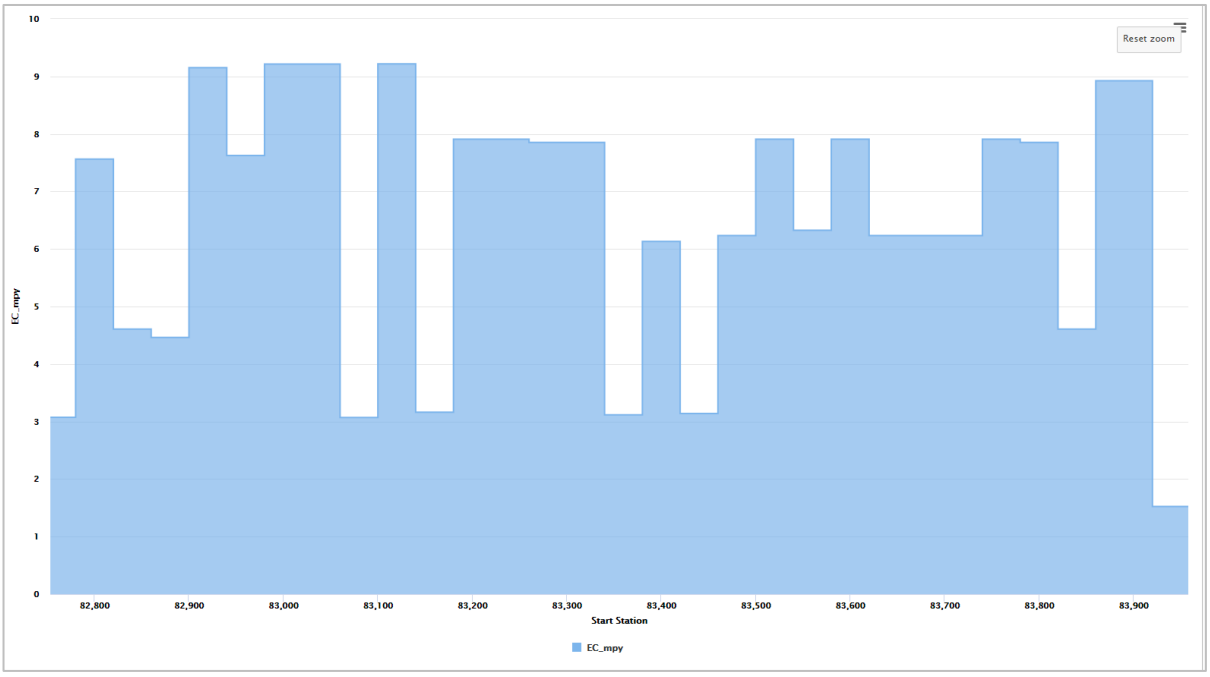


Figure 17 - mpy Results

Summary

Data driven machine learned classification and regression models and results provide a foundation for strategic risk based mitigative decision-making. Classification methods may be used to support susceptibility analysis and regression methods may be used to support severity analysis. The resulting learned models are explicitly validated with data in terms of confidence, accuracy and explainability. Data driven models provide a sound basis to support the pipeline industry's objective of zero unwanted events.

References

1. INGAA (2019) https://www.ingaa.org/Pipelines101/Safety.aspx
2. Brownlee, Jason (2013) "A Tour of Machine Learning Algorithms"
3. Saxena, Eklawerevya (2019) https://towardsdatascience.com/evaluating-a-machine-learning-algorithm-81746c947ad3
4. Reliability Engineering Resources (2018) https://www.weibull.com/