Practical Application of Machine Learning Methods to ILI Data

by Michael P. Gloven, PE
Managing Partner, EIS

Pipeline Pigging and Integrity Management Conference

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

Organized by
Clarion Technical Conferences
and **Tiratsoo Technical**

Overview

Machine learning is a field of computer science which uses statistical techniques and methods to give computer systems the ability to "learn" with data, without being explicitly programmed[1]. This paper demonstrates the application of several machine learning methods against in-line inspection defect data and its potential influencers. When applied in this context, these methods learn correlation and causation patterns which may then be applied to similar assets to predict or assess the presence, non-presence and severity of defects. The objective of the machine learning process is to provide an auditable, transparent and data-driven approach supporting asset management requirements.

Machine Learning Process

As shown in Figure 1, machine learning is a process which "learns" methods based on known observations of interest[2]. The learning methods, of which there are more than a hundred, are trained to become "models" or rule-sets which may then be applied to similar assets to predict outcomes.

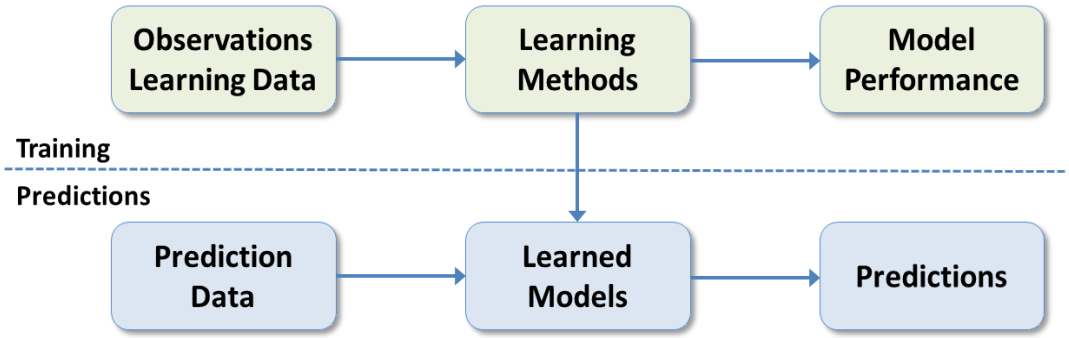


Figure 1 – Supervised Machine Learning Process

Most machine learning processes are referred to as supervised or unsupervised. A supervised process directs the method to a target of interest or "label" and may be used to predict the classification or severity of a defect. An unsupervised process is not directed towards a specific label and may be used to predict logical groups or clusters of data. These are the general types, although hybrids and other distinctions exist such as semi-supervised and reinforced learning.

An in-line inspection is expected to report defects which exceed specified thresholds along the pipeline asset. These reported defects are the "observations" and are combined with other potentially influencing data to learn supervised models.

As shown in Figure 2 and later in Table 1, learning methods may be grouped into different categories such as feature engineering, classification, regression, outlier detection, associations, clustering and time series[2]. The choice of category and method to use depends on many factors including business objectives, availability of data, expertise, and required transparency, complexity and performance.

Figure 2 – Machine Learning Algorithms & Methods[3]

Classification and regression methods are commonly used with in-line inspection data. Classification methods indicate patterns of higher or lower confidence of potential defects. These results might be used to support further investigation or inspection priorities. Regression when augmented with time series methods measure expected growth rates over time, the results which might be used to plan future assessments. Together, the learning results are deployable models supporting the assessment of defect likelihood and severity.

Machine learning is a process requiring involvement of experts in both machine learning methods and the subject matter area, ensuring the process, methods, data and results are valid and useful. Machine learning is a continuous improvement process as additional data is made available and models improve.

Benefits of Machine Learning

The machine learning process has the potential to deliver several benefits for the optimal use and analysis of in-line inspection data:

- ***Identification of High Susceptibility Areas*** - the application of resulting learned models identifies other areas with patterns of higher susceptibility for potential investigation or mitigation
- ***Optimal Assessment Planning*** – optimal inspection and assessment intervals may be established thru learned regressed time series analysis

- ***Transparent Validation & Performance*** **-** Uncertainty and accuracy are explicit properties of the models and results, informing the practitioner of data driven model reliability for decision-making

- ***Improved Data Management Strategies*** - data is explicitly measured in terms of importance and quality to the desired objective, which supports prioritization of data management processes

Machine Learning Building Blocks

Although the machine learning process can be performed on any in-line inspection reported defect, the remainder of this paper will focus on the building blocks or key elements to perform a machine learning project considering the in-line inspection feature "external corrosion metal loss" as the learning target and relevant design, operating and environmental influencers as the learning data. An important note regarding data, the term "feature" henceforth refers to a potentially influencing data field or data column and should not be confused with the feature designation within an in-line inspection. Feature is the term normally used in machine learning processes to describe data.

Data Preparation

Data preparation is a key element of the machine process and is generally the area requiring the most resources and effort. The objective of data preparation is to produce high quality and useful "Data Sets" for both learning and predictions. The learning data set includes the observations and features used to train or learn the selected method(s). The resulting model is applied to the prediction data set which generally excludes observation data but includes a broader set of similar assets.

Key processes of data preparation include initial data source selection by subject matter experts, data aggregation and dynamic segmentation, data quality measures and sampling. Both the learning and prediction data sets will use similar data preparation processes for each target of interest.

Numerous open source and purposed applications are available to support these processes[4]. Important considerations for application selection include the ability to persist, update and version processes, securely handle large amounts of data, and provide early visibility of data issues.

Data Aggregation & Dynamic Segmentation – The data aggregation and dynamic segmentation process identifies, aggregates and prepares data into a table of n-dimensional vectors, or more simply as data records comprised of at least one target and multiple features. As shown in Figure 3, each column of the external corrosion learning data set is a feature and the green shaded column is the target of interest.

For certain methods, the learning data set requires records for both true (defect present) and false (no defect present) observations. True observations are reported defects. False observations require a sampling strategy of units or pipe length since there often is no explicit false observation record other than no defect found above the reporting threshold. A common heuristic is to use a pipe joint length as the record length for a false observation (no defect found), although there are methods to optimize this assumption.

Row No.	EC	Bedrock	Corrosivity	CP_Off	CP_On	Diameter	DOC	Ext_Coating	Farmland	Flooding	Frost	Hwy_Type	Install_Yr	Power_Line	RAILROAD	Road_Type	Slope
14027	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14028	false	>40	High	-.800 to -.950	-.950 to -1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14029	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14030	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14031	false	>40	High	-.800 to -.950	<-1.100	12.750	>36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14032	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14033	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14034	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14035	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14036	false	>40	High	-.800 to -.950	-.950 to -1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14037	true	>40	High	>-.800	>-.850	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14038	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14039	false	>40	High	-.950 to -1.100	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14040	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14041	false	>40	High	-.800 to -.950	-.950 to -1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14042	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14043	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14044	false	>40	High	-.800 to -.950	-.950 to -1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14045	false	>40	High	-.800 to -.950	-.950 to -1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14046	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14047	false	>40	High	-.950 to -1.100	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14048	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14049	false	>40	High	-.950 to -1.100	-.950 to -1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14050	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20
14051	false	>40	High	-.800 to -.950	<-1.100	12.750	24 to 36	TGF	Farmland	No Data	Low	None	1974	None	None	None	>20

Figure 3 – Example Learning Data Set for External Corrosion

Data Quality Measures – Prior to performing any learning or predictions, data sets are validated by domain experts and those familiar with statistical techniques. The following quality measures are particular to the machine learning process and should be augmented with reviews of data accuracy, relevance and comprehensiveness against business objectives.

- Number of Learning Observations – A statistically significant number of observations should be present in the learning data set representing the population of assets and target(s). A common heuristic is to require 10-30 observations within the learning population for each target, however, the practitioner is encouraged to perform hypothesis testing to ensure adequate sample size for the target population.

- Data Stability – Stability is a measure of how data changes within the learning data set. If the data represents mostly one value, i.e. depth of cover is three feet for 99.9% of the data, the learning from this data may be less than optimal.

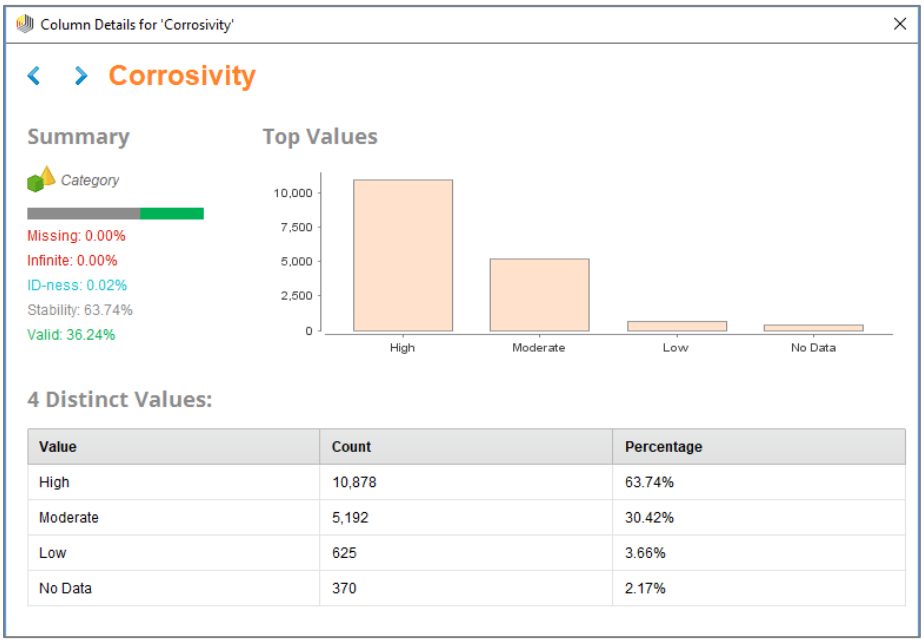
- Data Attribute Variance – Attribute variance is a measure of the number of attributes representing the data i.e. if there are 1000 coating types in the learning data, the learning from this data may be less than optimal versus if there are only 10 types.

- Data Id-ness – Id-ness is a measure of how data represents a unique id and does not vary. Data that acts like an id may not provide useful information for learning.

- Data Missing – Missing data is problematic for many machine learning methods since a missing value may be similar to a null and methods such as regression are unable to calculate an n-dimensional point in space if a null value is included in its vector coordinate. Missing data may be resolved thru inference, imputing, SME opinion or data collection efforts.

- Temporal State – Temporal state measures the relevance of data based on its shelf life and usefulness to the analysis. Temporal measures are normally performed by SME's who have an understanding of data correlation and causation. Older data may be deemed irrelevant and removed from the data sets.

These measures augmented with reviews of accuracy, relevance and comprehensiveness may be assessed thru different scoring approaches such as a high, medium, low or 0-10 scale, or simply go-no-go criteria for potential mitigation. Figure 4 is a sample quality measure report for soil corrosivity. It's important to consider and mitigate these measures prior to moving forward with learning and prediction processes.



Value	Count	Percentage
High	10,878	63.74%
Moderate	5,192	30.42%
Low	625	3.66%
No Data	370	2.17%

Figure 4 – Quality Measure Report

Sampling – Adequate sample size and sampling is used in the learning process to train and validate models, but is a consideration in the data preparation process as this is where records originate. In its basic form, sampling holds back a percentage of observations to validate the trained model. Cross validation based sampling is a popular structure for iterating thru the learning data set in multiple layers and averaging the learned models into one final model. Other methods include split validation, up-sampling, stratified and class balancing. Machine learning requires a strategic approach to sampling to support optimal learning.

Other Data Preparation Processes – Different machine learning methods may require additional data preparation processes. For example, regression can only work with numerical data, therefore, categorical data must be converted to binominal 0/1 or equivalent values. K-nearest neighbour methods generally require normalization of numerical data so large numbers do not bias the distance calculations. Decision trees can accept most n-dimensional data and are often the first method used to gain initial insights into predictive patterns. The practitioner should research and understand the limitations and constraints of any method they decide to deploy for learning.

Upon completion of the data preparation process, learning and prediction data sets have been created and a sampling method designed for learning external corrosion metal loss. The next step in this iterative process is to understand data features in greater detail for potential mitigative action.

Feature Engineering and Selection

Feature engineering is a collection of methods to measure the importance of individual features to the particular target or objective. Feature engineering is used to optimize the number of features for the analysis and addresses the paradox of Occam's Razor which states that "simpler solutions are more likely to be correct than complex ones".

Correlation vs. Causation

Machine learning reveals underlying patterns comprised of features. These features may be further characterized as correlation or causation. An interpretation in the domain of pipeline defects might be that correlations are coincidental associations whereas causations are features which may be altered in the future to have an impact on the predicted defect. Often, it is the subject matter domain expert making these interpretations and initial selection of causation features.

Numerous methods are available for feature engineering, the following are common with practitioners. The objective is to understand the data and determine if it's sufficient and useful to the learning process.

Deviation Analysis

Deviation analysis is useful for determining the impact of specific features on the target of interest, in this case external corrosion. Figure 5 shows each feature in the learning data set and its mean and standard deviation value for when external corrosion is true (red) or false (blue). Large separations in the means indicate the ability of the feature to classify the presence of external corrosion and the shaded areas indicate the spread of underlying data. Small separations in means indicates minimal

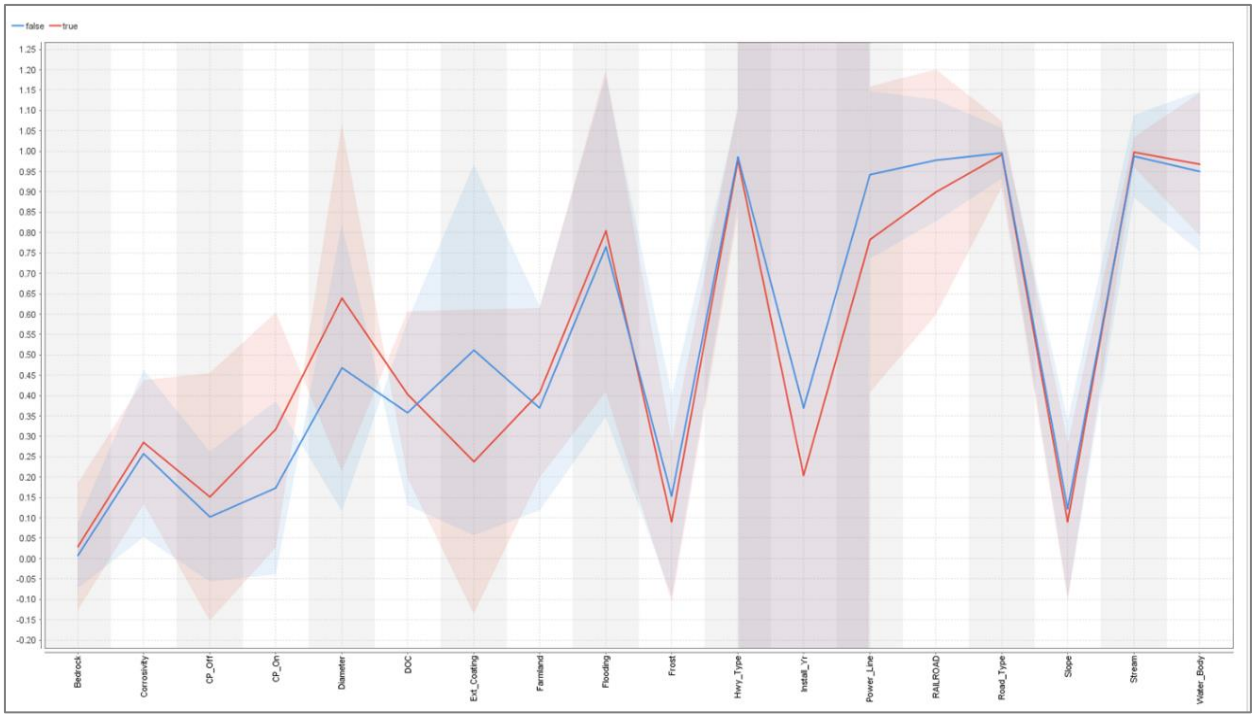


Figure 5 – Deviation Analysis External Corrosion Features

ability to classify. The largest normalized mean deviation is "external coating" indicating its potential influence to external corrosion. This analysis is a useful two-dimensional first cut approach to evaluating potential causation learning data.

Correlation Analysis

Correlation coefficients and its square, coefficient of determination, are useful in evaluating the relationship of features to features and features to targets. A positive correlation coefficient of 1 indicates a direct positive relationship between features whereas a -1 indicates a negative correlation. The coefficient of determination will be used later in measuring regression performance, however, it can be used here within feature engineering to measure the explained variation over total variation. A high coefficient of determination indicates the underlying independent variable or feature does well explaining the result. As shown in Figure 6, a sampling based correlation method indicates cp readings correlate the best against external metal loss, although only 24% of the relationship is explained and 76% unexplained.

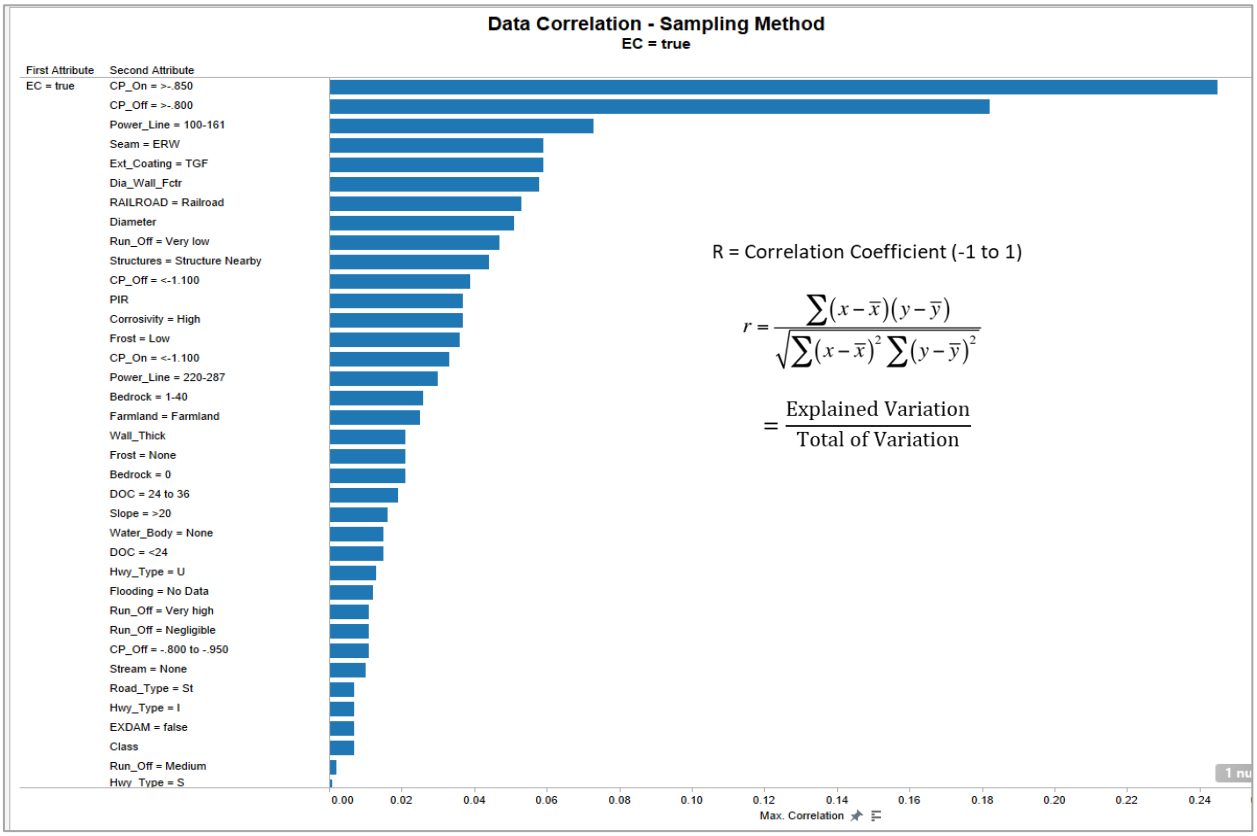


Figure 6 – Correlation Coefficient

Correlation values support data management strategies in several areas. For example, highly correlated values for soil corrosivity and soil type may suggest the analysis does not require both features in the data sets since their influence is the same. At the same time, a benefit of highly correlated features is if a value is missing from a record, the other value can be used to infer or impute feature data. Conversely, some methods such as regression will perform poorly if highly correlated data is not removed i.e. only retain one representative feature. As with deviation analysis, correlation is another method to understand data underlying the observation and prepare the data for learning.

Entropy & Information Gain

Feature entropy calculations and information gain is a method to determine the ability of a feature to classify a target. Shannon's Entropy[5] may be calculated for each feature and target where p equals the distribution of the feature for each target class (Figure 7). The feature which lowers the entropy the most from the target entropy value is normally the most influential feature in classifying the target and hence has the largest information gain value.

Shannon's Entropy

$$H = -p*\log_2(p) - p*\log_2(p)$$

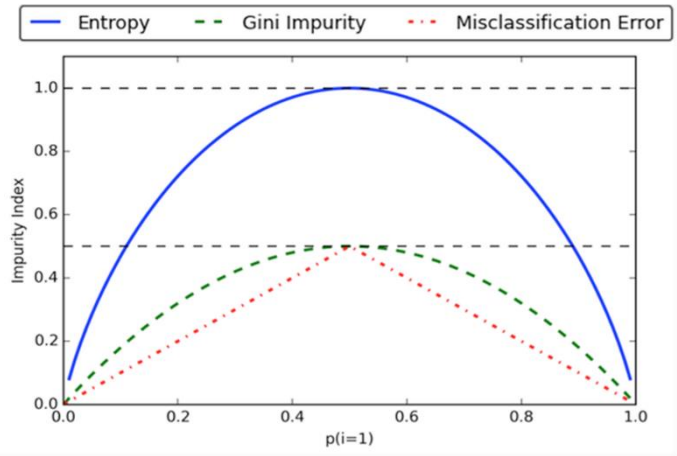


Figure 7 – Entropy & Gini Index

Figure 8 provides an output of normalized information gain ratio calculations for learning feature data against external corrosion. Applying Shannon's Entropy to each feature, taking the difference in entropy from the target data distribution and normalizing the results to 1 indicates the presence of a railroad is most influential to classifying the presence or non-presence of external corrosion. On the other hand, slope of the ground has the least influence on the presence of external corrosion. As a

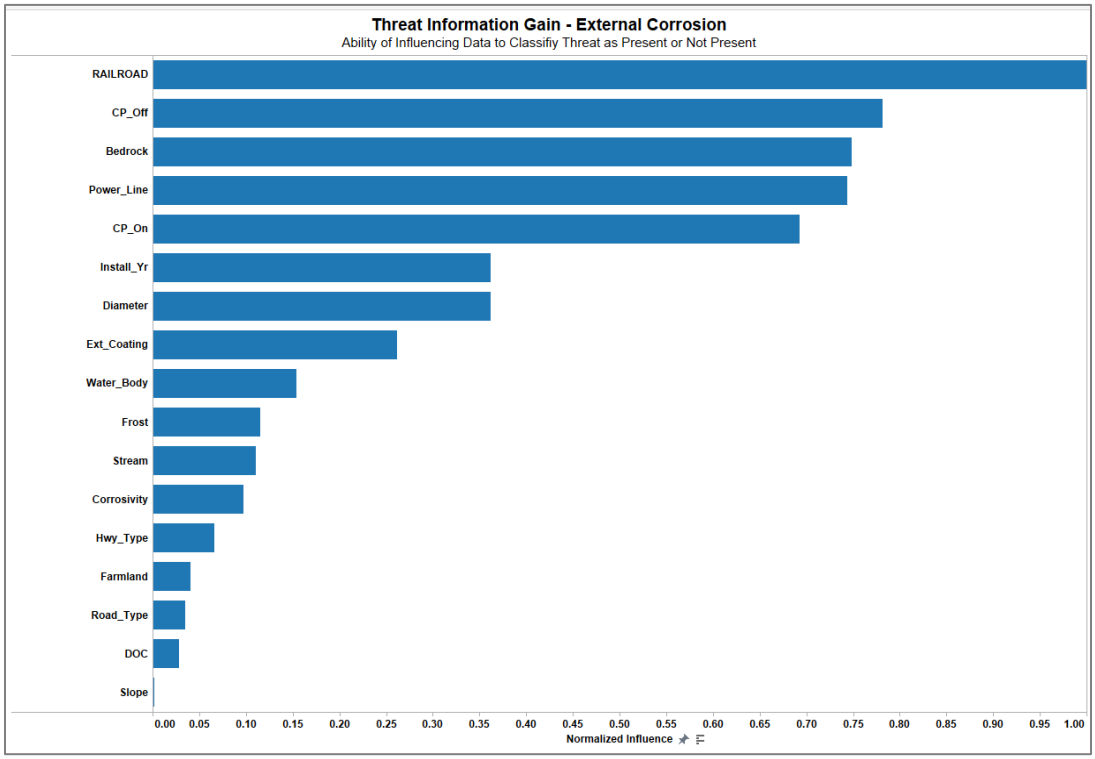


Figure 8 – Feature Information Gain Ratio

reminder, information gain is performed in the context of the actual observation of external corrosion defects.

Evolutionary Optimization

An emerging method for feature engineering is evolutionary optimization, a genetic algorithm (GA) search heuristic that mimics the process of natural evolution[6]. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

Figure 9 shows an output of GA as a pareto front optimizing a collection of features along an accuracy scale. This example indicates two features are capable of an 82% prediction accuracy whereas increasing to twelve features provides an accuracy of 97%. GA supports the analysis of trade-offs between feature selection simplicity and complexity, and required performance.

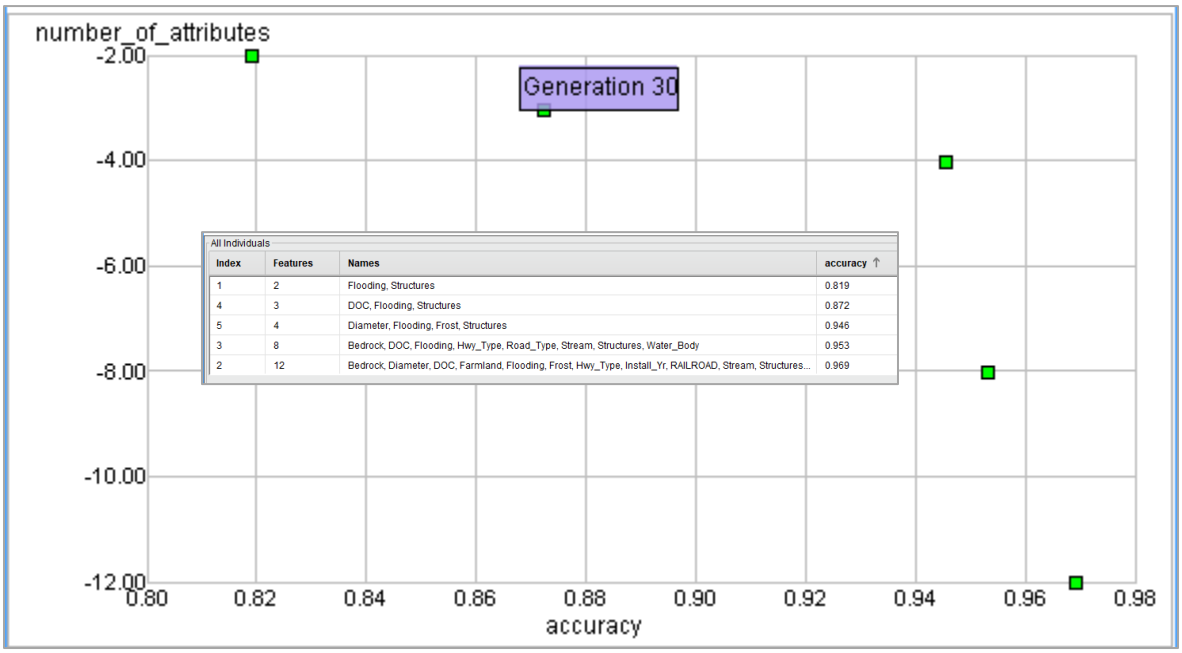


Figure 9 – GA Pareto Front Feature Selection

In terms of the external corrosion example, GA selects an optimal number of features required to meet a specific performance target. For example, if there are 100 features potentially influencing external corrosion and are seeking a 90% accurate model, the approach would find the best combination of features to meet this performance value. The end result might be a small percentage of the features (less than 100) which would then help simplify the learning process and cost of data collection.

Benefits of Feature Engineering

For each of the presented methods, the practitioner is learning about the data prior to further use and this learning may result in specific strategic actions. For example, if each method shows poor deviation, no correlation, limited information gain and an inadequate combination of data to meet performance requirements, it is likely that there is not enough features to provide a useful external learning model. The action may then be to collect additional features and possibly remove useless

features. On the other hand, good deviations, correlations, information gain and genetic selection indicate useful models are possible with the current learning data.

It is useful to note that each method provides some variation of insights and priority into the data as it applies to the target, in this case, external corrosion metal loss. The practitioner should have a solid understanding of the math and reasons for these variations prior to making strategic data decisions.

At this point in the process, important features within the data sets have been identified in preparation for the next step, method learning.

Learning Methods & Models

The main idea behind complex systems is that the ensemble behaves in ways not predicted by its components as the interactions matter more than the nature of the units[7]. In other words, except for the GA method, the previous feature engineering related methods are useful in understanding one-to-one relationships with data, however, it's the interaction or many-to-many relationships which may be more important in learning underlying patterns. This is an area where machine learning methods excel and become most valuable. As shown in Table 1, numerous methods are available to the practitioner to learn external corrosion:

Categories	Description	Methods	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Determining the probability of the presence of external corrosion
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from known data set	Linear regression, Logistic regression	Determining external corrosion growth rates
Outlier Detection	Predict if a data point is an outlier compared to other data points in the data set	Distance based, Density based, LOF	Identifying potential rare event occurrences
Time series	Predict if the value of the target variable for future time frame based on historical values	Weibull, Exponential smoothing, ARIMA	Identifying external corrosion probabilities over time
Clustering	Identify natural clusters within the data set based on inherit properties within the data set	K means, density based clustering - DBSCAN	Identify logical groupings of data exclusive of external corrosion, determine if data organizes as expected

Table 1 – Categories of Learning Methods

Each method is learned through the learning data set, the output being a model which can then be applied to similar assets. For a comprehensive understanding of the algorithms and math behind the methods, the practitioner is encouraged to study Stanford's on-line training course in machine learning led by Andrew Ng[8]. Numerous free on-line training resources are available which are supported through open source and purposed machine learning applications. Next is a discussion of common methods in the context of external corrosion metal loss.

Classification

Classification methods are useful in determining the confidence an observation of interest will be true or false. Figure 8 shows seven classification methods applied against the same data sets and the variation in resulting patterns. Note the different separations shown as blue and red of this binominal classification problem (i.e. the presence or non-presence of external corrosion). The methods produce varying patterns with the level of shading indicating confidence of the model results.

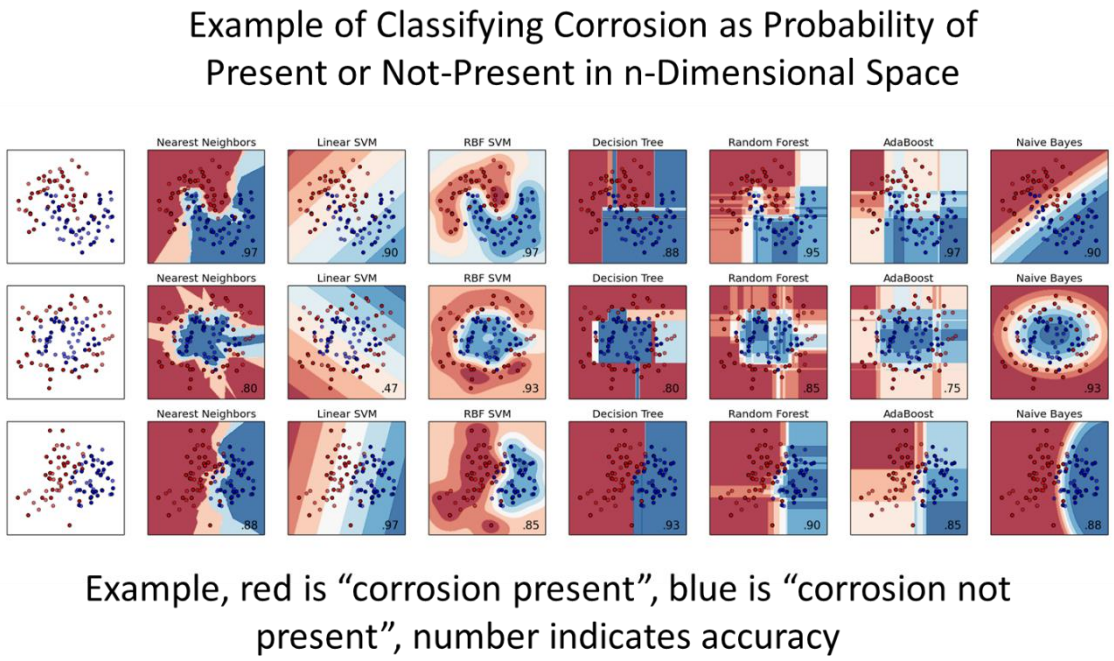


Figure 10 – Classification Examples

Classification methods provide a confidence value that a prediction will be classified as true\false or polynominal value. Decision tree, bayesian, nearest neighbour and GBT classifiers are most popular as they are easy to interpret and understand. Deep learning methods are also popular but require expertise in the underlying algorithms to understand how they work. Figure 11 shows a learned decision tree model as expressed as a rule-set and Figure 12 shows the details of a learned deep learning model. Each model uses the same learning data set.

Tree

```
CP_On = -.850 to -.950
|   Frost = High
|   |   Flooding = Frequent: false {false=87, true=0}
|   |   Flooding = No Data: true {false=0, true=4}
|   |   Flooding = None: true {false=0, true=6}
|   |   Flooding = Occasional
|   |   |   Diameter > 8.610: false {false=173, true=12}
|   |   |   Diameter ≤ 8.610: true {false=0, true=6}
|   Frost = Low
|   |   Ext_Coating = FBE
|   |   |   Corrosivity = High: true {false=0, true=60}
|   |   |   Corrosivity = Moderate: false {false=3, true=0}
|   |   |   Corrosivity = No Data: true {false=0, true=2}
|   |   Ext_Coating = PE
|   |   |   Power_Line = 100-161: true {false=22, true=157}
|   |   |   Power_Line = 220-287: true {false=1, true=16}
|   |   |   Power_Line = None: true {false=233, true=487}
|   |   Ext_Coating = TGF: false {false=12, true=0}
|   Frost = Moderate
|   |   Corrosivity = High
|   |   |   Flooding = Frequent: false {false=77, true=0}
|   |   |   Flooding = None: false {false=330, true=83}
|   |   |   Flooding = Occasional: true {false=0, true=15}
|   |   Corrosivity = Low: false {false=192, true=0}
|   |   Corrosivity = Moderate: false {false=38, true=0}
|   |   Corrosivity = No Data: true {false=0, true=2}
|   Frost = No Data: false {false=30, true=0}
|   Frost = None
|   |   Bedrock = 0
|   |   |   Stream = None: true {false=86, true=141}
|   |   |   Stream = Stream/River: true {false=10, true=60}
|   |   Bedrock = >40: true {false=0, true=32}
CP_On = -.950 to -1.100
|   Corrosivity = High
|   |   Install_Yr > 1991.500
|   |   |   Frost = High: true {false=0, true=2}
|   |   |   Frost = Low: false {false=694, true=64}
|   |   |   Frost = Moderate: false {false=1138, true=27}
|   |   |   Frost = No Data: false {false=16, true=0}
|   |   |   Frost = None: true {false=0, true=23}
|   |   Install_Yr ≤ 1991.500
|   |   |   Install_Yr > 1981.500: true {false=0, true=732}
|   |   |   Install_Yr ≤ 1981.500: false {false=4266, true=4206}
|   Corrosivity = Low
```

Figure 11 – Decision Tree Model

DeepLearning

```
Model Metrics Type: Binomial
 Description: Metrics reported on temporary training frame with 9951 samples
 model id: rm-h2o-model-deep_learning_(2)-187820
 frame id: rm-h2o-frame-deep_learning_(2)-759831.temporary.sample.29.63%
 MSE: 0.3837978
 R^2: -0.5354035
 AUC: 0.8490602
 logloss: 1.9856867
 CM: Confusion Matrix (vertical: actual; across: predicted):
        false  true   Error            Rate
 false   3596  1438  0.2857  = 1,438 / 5,034
  true    845  4072  0.1719  =   845 / 4,917
Totals   4441  5510  0.2294  = 2,283 / 9,951
Gains/Lift Table (Avg response rate: 49.41 %):
  Group  Cumulative Data Fraction  Lower Threshold      Lift  Cumulative Lift  Response Rate  Cumulative Response Rate  Capture Rate  Cumulative Capture Rate        Gain  Cumulative Gain
      1                0.01004924         0.983043  2.023795        2.023795       1.000000                  1.000000      0.020338                 0.020338  102.379500       102.379500
      2                0.02220882         0.928509  2.023795        2.023795       1.000000                  1.000000      0.024609                 0.044946  102.379500       102.379500
      3                0.03225806         0.802397  1.922605        1.992272       0.950000                  0.984424      0.019321                 0.064267   92.260525        99.227171
      4                0.04009647         0.680470  1.686496        1.932496       0.833333                  0.954887      0.013219                 0.077486   68.649583        93.249597
      5                0.05376344         0.661760  1.711297        1.876266       0.845588                  0.927103      0.023388                 0.100875   71.129724        87.626602
      6                0.10491408         0.355932  1.856802        1.866776       0.917485                  0.922414      0.094977                 0.195851   85.680209        86.677642
      7                0.15284896         0.150743  1.930454        1.886746       0.953878                  0.932281      0.092536                 0.288387   93.045435        88.674642
      8                0.20008039         0.083585  1.761132        1.857094       0.870213                  0.917629      0.083181                 0.371568   76.113224        85.709365
      9                0.30620038         0.020773  1.458436        1.718931       0.720644                  0.849360      0.154769                 0.526337   45.843560        71.893057
     10                0.40649181         0.011183  1.222794        1.596522       0.604208                  0.788875      0.122636                 0.648973   22.279397        59.652159
     11                0.52155562         0.002861  1.203672        1.509852       0.594760                  0.746050      0.138499                 0.787472   20.367196        50.985245
     12                0.60024118         0.000789  0.904634        1.430514       0.446999                  0.706847      0.071182                 0.858654   -9.536622        43.051439
     13                0.70013064         0.000094  0.853089        1.348132       0.421529                  0.666140      0.085215                 0.943868  -14.691136        34.813156
     14                0.80554718         0.000001  0.449518        1.230536       0.222116                  0.608034      0.047387                 0.991255  -55.048214        23.053603
     15                0.90704452         0.000000  0.084158        1.102258       0.041584                  0.544649      0.008542                 0.999797  -91.584219        10.225750
     16                1.00000000         0.000000  0.002188        1.000000       0.001081                  0.494121      0.000203                 1.000000  -99.781211         0.000000
Status of Neuron Layers (predicting EC, 2-class classification, bernoulli distribution, CrossEntropy loss, 6,452 weights/biases, 81.4 KB, 337,520 training samples, mini-batch size 1):
 Layer Units      Type Dropout       L1       L2 Mean Rate Rate RMS Momentum Mean Weight Weight RMS Mean Bias Bias RMS
     1    75     Input  0.00 %
     2    50 Rectifier  0.00 % 0.000010 0.000000  0.267840 0.430786 0.000000   -0.012839   0.158811  0.391197 0.110897
     3    50 Rectifier  0.00 % 0.000010 0.000000  0.037929 0.134827 0.000000   -0.049210   0.225761  0.634527 0.272640
     4     2   Softmax         0.000010 0.000000  0.004988 0.005038 0.000000    0.020921   0.439642  0.013343 0.110902
Scoring History:
            Timestamp   Duration  Training Speed   Epochs Iterations       Samples Training MSE Training R^2 Training LogLoss Training AUC Training Lift Training Classification Error
 2019-01-07 10:45:09  0.000 sec                  0.00000          0      0.000000           NaN          NaN             NaN          NaN           NaN                         NaN
 2019-01-07 10:45:10  0.388 sec 107833 rows/sec  1.00000          1  33752.000000       0.33899     -0.35617         2.69490      0.74583       1.52425                     0.33092
 2019-01-07 10:45:12  2.285 sec 155467 rows/sec 10.00000         10 337520.000000       0.38380     -0.53540         1.98569      0.84906       2.02379                     0.22942

H2O version: 3.8.2.6-rm9.0.0
```

Figure 12 – Deep Learning Model

Regression

Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable (i.e. the label) and a series of other changing variables known as independent variables (features). Just like Classification is used for predicting categorical labels, regression is used for predicting a continuous dependent value. Linear regression attempts to model the relationship between a scalar variable and one or more explanatory variables by fitting a linear equation to observed data.

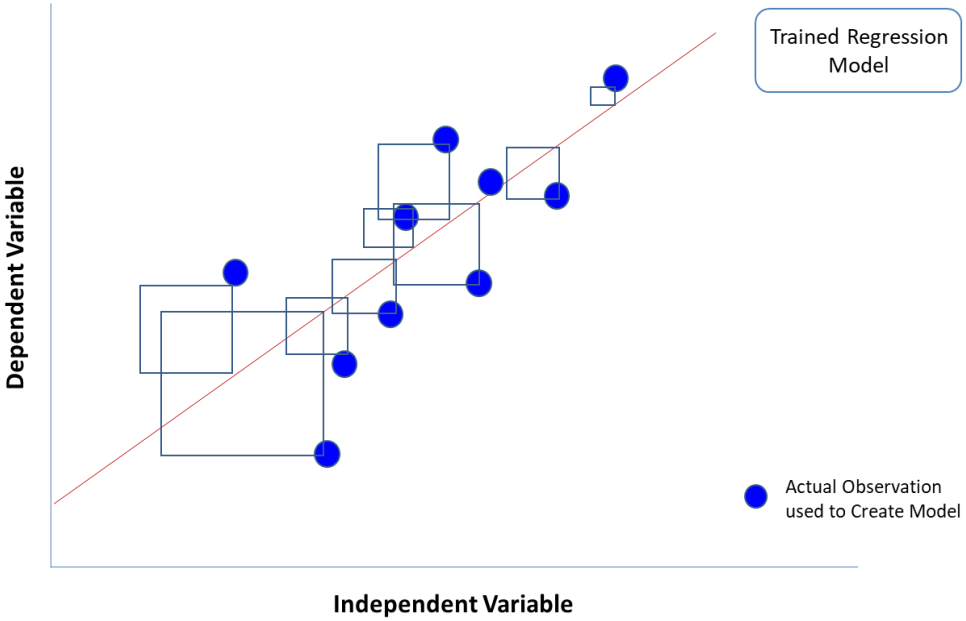


Figure 13 – Linear Regression OLS Concept

A popular regression method is multi-variate linear regression using ordinary least squares to fit the prediction line. Figure 13 shows the concept where the method attempts to find the best fit based on OLS.

The actual learned model thru linear regression may be represented thru Figure 14. The model coefficients are multiplied by the attribute value to predict the dependent variable, in this case, external corrosion wall loss. The additional columns are used to assess the significance of each feature.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Bedrock = >40	1.546	0.041	3.440	0.998	38.172	0	****
Bedrock = 1-40	1.614	0.061	2.377	0.999	26.387	0	****
Bedrock = 0	1.127	0.053	1.910	0.999	21.207	0	****
Corrosivity = Moderate	1.080	0.020	4.829	1.000	53.626	0	****
Corrosivity = High	1.079	0.019	5.045	0.999	56.006	0	****
Corrosivity = Low	0.972	0.078	1.172	0.921	12.490	0	****
Corrosivity = No Data	1.156	0.075	1.393	0.995	15.439	0	****
CP_Off = -.800 to -.950	1.031	0.019	5.288	0.900	55.705	0	****
CP_Off = -.950 to -1.100	1.044	0.022	4.306	0.992	47.635	0	****
CP_Off = <-1.100	1.055	0.054	1.790	0.977	19.648	0	****
CP_Off = >-.800	1.156	0.036	3.897	0.556	32.257	0	****
DOC = >36	-0.013	0.020	-0.060	0.950	-0.652	0.516	
DOC = 24 to 36	0.014	0.019	0.067	0.946	0.724	0.470	
Ext_Coating = TGF	1.404	0.019	7.732	0.746	74.162	0	****
Ext_Coating = PE	1.393	0.021	6.154	0.910	65.195	0	****
Ext_Coating = FBE	1.490	0.023	6.328	0.883	66.056	0	****
Farmland = Not Farmland	1.442	0.019	6.743	0.999	74.854	0	****
Farmland = Farmland	1.477	0.019	7.041	1.000	78.191	0	****
Farmland = No Data	1.368	0.075	1.649	0.996	18.274	0	****
Flooding = None	0.987	0.019	4.565	1.000	50.693	0	****
Flooding = Frequent	0.939	0.106	0.803	0.989	8.865	0.000	****
Flooding = Occasional	1.403	0.110	1.199	0.914	12.736	0	****
Flooding = No Data	0.958	0.019	4.524	0.993	50.063	0	****

Figure 14 – Linear Regression Model

Time Series

Time series methods consider the relationship of time and the dependent variable or target. In general but not always, classification and non-temporal regression methods determine models for a

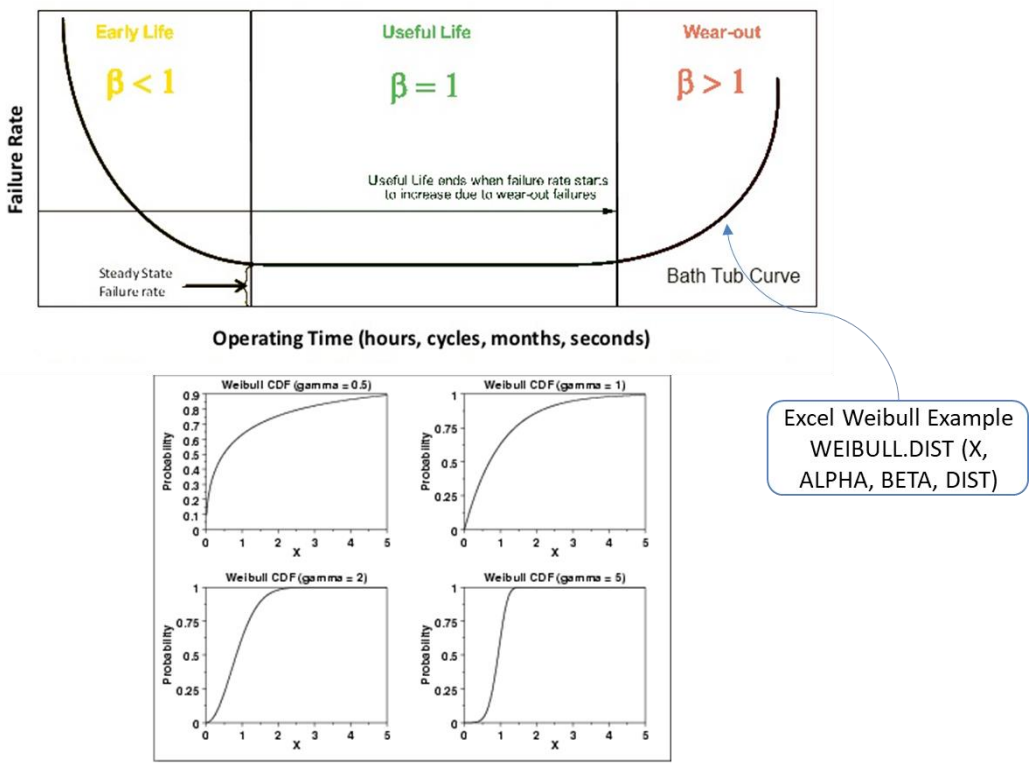


Figure 15 – Reliability & Survival Curves

point in time, not looking backwards or forwards. Time series methods allow for more explicit predictions at different points in time.

Common methods include ARIMA (Autoregressive Integrated Moving Average) and 2 or 3 parameter Weibull[9] curves. Figure 15 shows a 2 parameter Weibull curve for prediction of time dependent defects in the wear-out portion of the bathtub curve. The parameters are either provided by the originators of the asset or calculated thru Benard's approximation or equivalent means.

Outlier Detection

Outlier detection is a collection of methods to find anomalous multi-dimensional points of interest. A common method is to set a standard deviation range to find anomalies. More sophisticated methods are distance or outlier factor based on k^{th} nearest neighbour and Breunig algorithms. The purpose of outlier detection is to potentially find and understand rare events or anomalous conditions.

Clustering

Clustering is a collection of unsupervised methods to reveal logical groupings or organizations of data in n-dimensional space. Figure 16 illustrates thru colours a cluster of data.

There are numerous clustering methods, popular approaches being x-means and k-nearest neighbour. Figure 17 shows an x-means based model for external corrosion indicating four optimal clusters. The results may be used to substantiate beliefs or bias of data. Note how the data results are numerical and represent the centroid of each cluster.

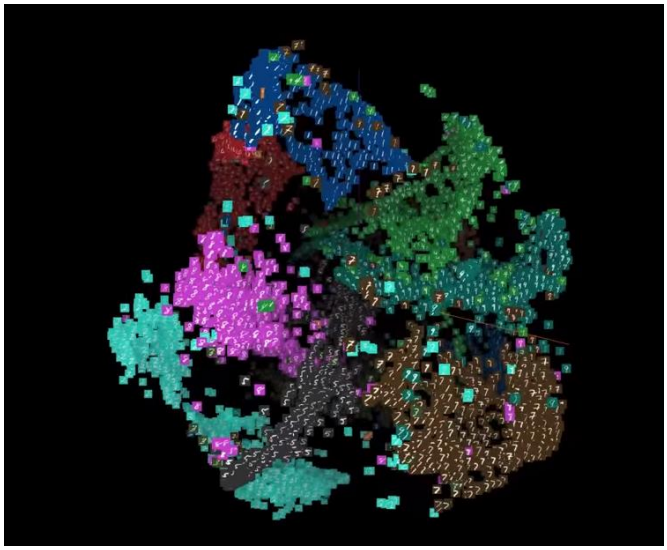


Figure 16 – Cluster Illustration

Attribute	cluster_0	cluster_1	cluster_2
Bedrock	0.073	0	0
Corrosivity	1.014	0.468	0.796
CP_Off	0.510	0.444	0.173
CP_On	0.577	0.125	0.840
Diameter	8.625	6.625	14.050
DOC	0.564	0.281	0.932
Ext_Coating	2	2	0
Farmland	0.852	0.562	0.753
Flooding	2.048	4	3.691
Frost	0.880	0.350	0.661
Hwy_Type	2.980	2.995	2.941
Install_Yr	1992	1981	1973.200
Power_Line	1.815	1.951	1.859
RAILROAD	0.991	1	0.968

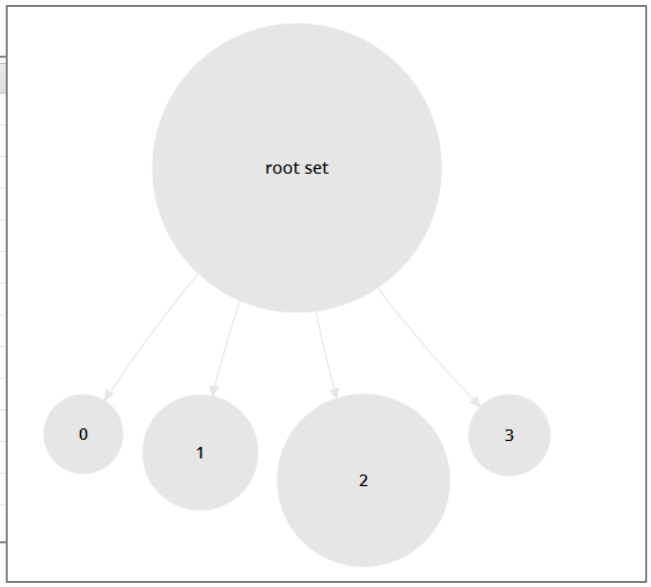


Figure 17 – Cluster Model for External Corrosion

Model Performance

Selecting the appropriate method or model is not an easy task. There are numerous methods for learning, and a good practice is to select the best method using common criteria. Suggested criteria are model transparency, cost to implement, time to learn and execute, complexity and performance.

The following outlines the main concepts in measuring accuracy, or as referred to in machine learning, the performance vector. Content and documentation regarding performance measurement is extensive. Common approaches to measure and understand performance include the confusion matrix, ROC curve, RMSE, R^2 and bias\variance.

Confusion Matrix

The confusion matrix is used to measure the performance of classification models. Recall that held-back data or actual observations are used to validate and test the resulting model. The confusion matrix provides the results of validation as shown in Figure 18.

The matrix indicates how well the model performs against the test data. Note the measures of false positives and misses. A strategy will be required regarding acceptable performance. For example, maybe sensitivity is more important to the analysis, that is, not missing actual defects. Or maybe there needs to be a balance between false positives and misses, with accuracy being a secondary measure. This is an important consideration in method selection as each of the learned models would be expected to have a variation of the performance vector characteristics.

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

The performance vector generates four primary measures:

- Accuracy = (TP + TN)/(TP + TN + FP + FN)
- Precision = TP/(TP + FP)
- Sensitivity = TP/(TP + FN)
- Specificity = TN/(TN + FP)

Figure 18 – Confusion Matrix

ROC Curves

As shown in Figure 19, Receiver Operator Characteristic (ROC) curves illustrate the rate of true positives (y-axis) vs. the rate of false positives (x-axis) for different methods. The higher to the left the method, the better the performance (AUC Area Under Curve approaches 1).

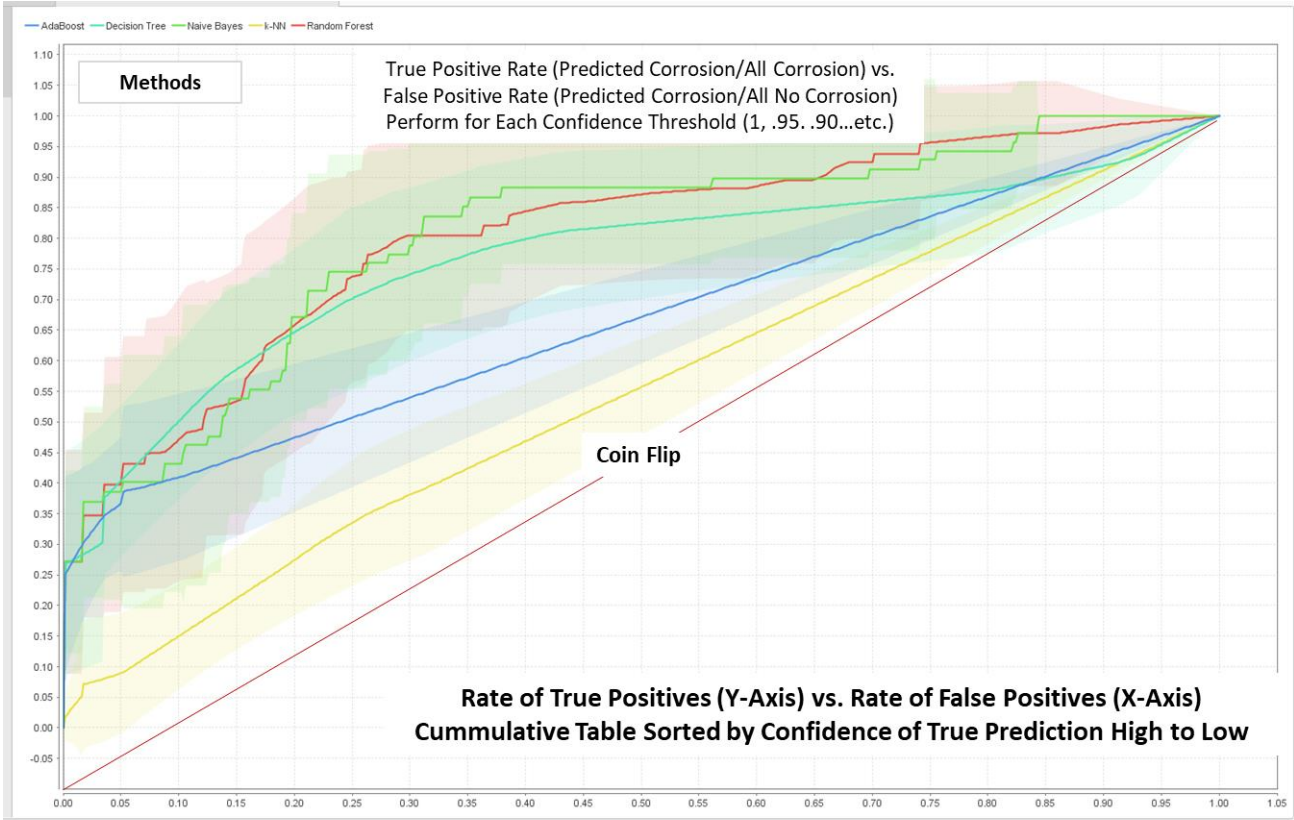


Figure 19 – ROC Curves

Regression Performance

Regression performance can be measured thru several methods, the most common being Root Mean Squared Error (RMSE) and Squared Correlation (R^2). RMSE is simply the square root/(n-1) of the boxed areas as shown in Figure 13. R^2 is a concept introduced in feature engineering and is a measure of how well the independent variables explain the change in dependent variable. Models with R^2 approaching 1 indicate variations are explained well, although this is somewhat subjective since the pipeline industry has not yet established acceptable correlation criteria.

Bias and Variance

Under and over-fitting are key concepts to understand in regression and classification performance, and best illustrated thru regression in Figure 20. Models which perfectly fit the training data may be "overfit" and thus when applied to prediction data may output high variance results. Models which are more generalized with bias may be "underfit" and may output high bias results. An objective of the practitioner is to find a balance between bias and variance, or complexity and simplicity in the learning process. This can be accomplished thru sampling, validation approaches, model selection and optimization of method specific parameters.

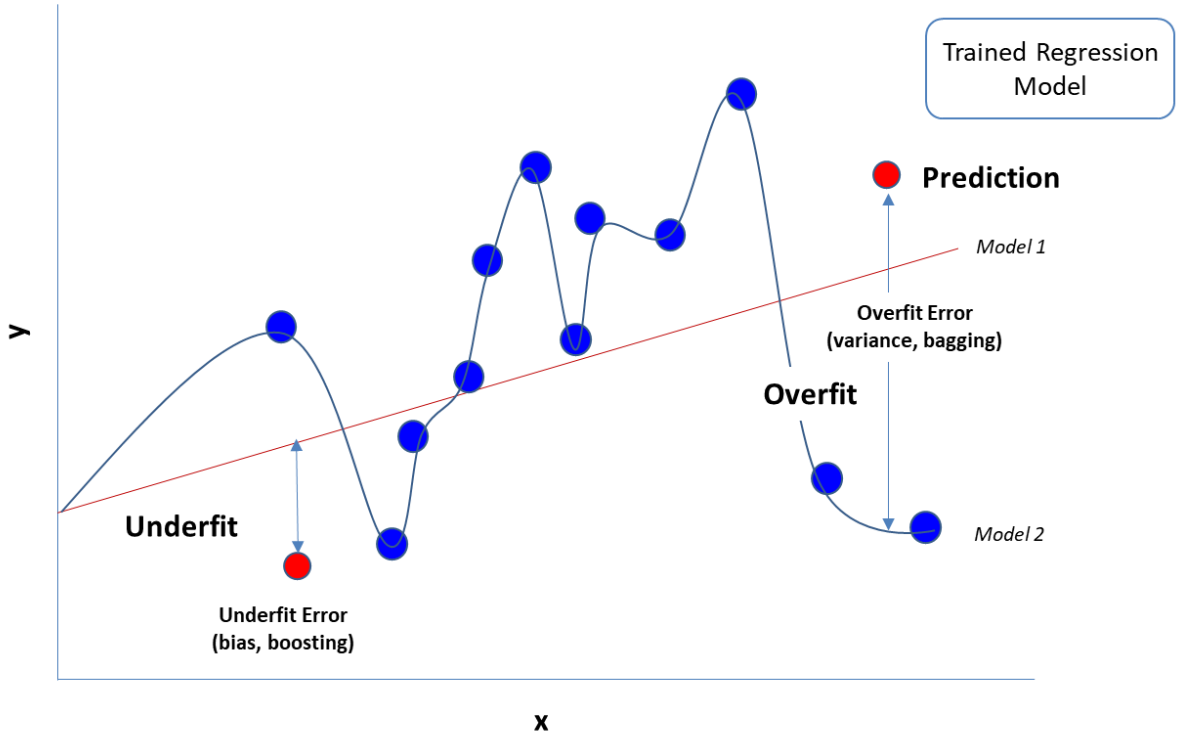


Figure 20 – Bias and Variance

Model Application and Results

Referring to Figure 1, the final step in the machine learning process is application of the selected learned model to the assets where prediction analysis is desired. The models are learned from the in-line inspection metal loss defect data and applied to assets of similar type to determine levels of susceptibility or severity of defects. The prediction data set has been processed similar to the learning data set, therefore, application is straight-forward since the feature data aligns with the learning model meta-data (features).

Results of model application are useful to show in a GIS or line graph format. Figure 21 shows machine learned predictions in a line graph, the top view being multivariate linear regressed results and the bottom view deep learning classification or susceptibility results. In this case, each dynamic segment will have a learning result based on the applied learned model and underlying predictor data.

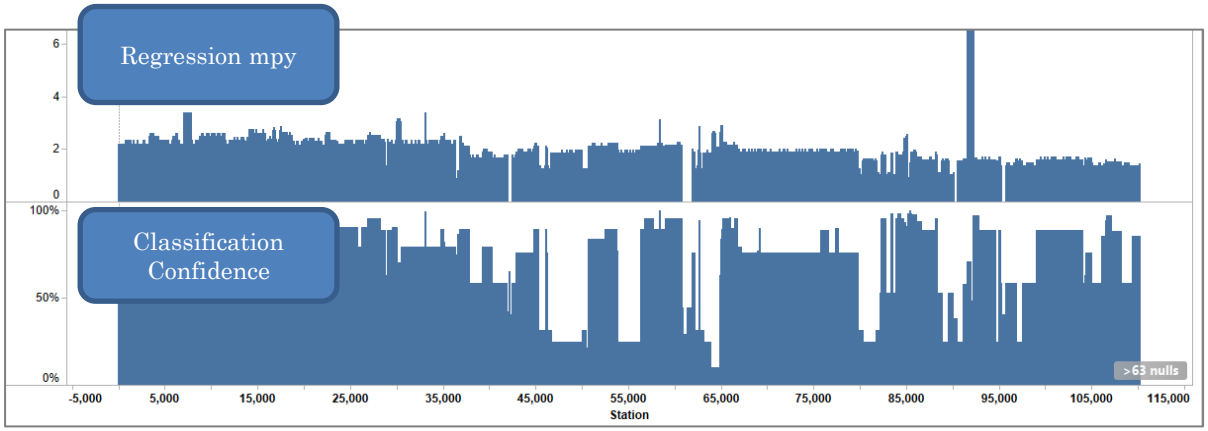


Figure 21 – Machine Learning Results

The practitioner will note that the results are associated with an explicit performance vector. A strategic question to consider in the analysis is what level of performance of the model and results triggers action. For example, is it best to focus action on an area with a 50% accurate model generating a classification confidence >75%, or a 75% accurate model generating a classification confidence >50%. Further, what level of prediction should drive what type of action. These are strategic questions beyond the scope of machine learning process fundamentals.

Lastly, as with machine learning methods, there are numerous applications to help visualize and analyse the results. Since the process is data driven, it's helpful to have an application which can view results top-bottom-top to gain a comprehensive understanding of results for the target of interest.

Summary

Machine learning is a valuable process to augment the analysis of in-line inspection results. The process and methods evaluate data quality and learn patterns to predict and assess the presence, non-presence and severity of defects over time which then provides useful knowledge to support asset management objectives.

References

1. Samuel, Arthur L. (1959). John McCarthy; Edward Feigenbaum (1990). "In Memoriam Arthur Samuel: Pioneer in Machine Learning"
2. Kotu, Vijay; Deshpande, Bala (2015) “Predictive Analytics and Data Mining”
3. Brownlee, Jason (2013) “A Tour of Machine Learning Algorithms”
4. KDnuggets (2018) “What software you used for Analytics, Data Mining, Data Science, Machine Learning projects in the past 12 months?” https://www.kdnuggets.com/2018/05/new-poll-software-analytics-data-mining-data-science-machine-learning.html
5. Shannon, Claude (1948) "A Mathematical Theory of Communication"
6. RapidMiner (2018) https://rapidminer.com/resource/webinar-better-machine-learning-models-multi-objective-optimization/
7. Taleb, Nicholas (2018) “Skin in the Game”
8. Coursera (2018) Courses in Machine Learning, https://www.coursera.org/learn/machine-learning
9. Reliability Engineering Resources (2018) https://www.weibull.com/