FRACTA

MACHINE LEARNING FOR WATER MAIN PIPE CONDITION ASSESSMENT

Prepared by: **Irene A. Onyeneho**, PhD,
Product Manager Machine Learning,
Data Engineering at Fracta.ai

1. INTRODUCTION

Machine Learning (ML) and Artificial Intelligence (AI) are two terms that are becoming more frequently used within our current technological ecosystem. Despite their growing prominence, few understand what these terms mean and how they can be applied to address significant business problems and direct strategy.

Market intelligence research firm IDC expects AI driven systems will drive worldwide revenues to over $47B by the year 2020, up from $8B in 2016.

1-1. WHAT IS ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING?

Artificial Intelligence is the science, technology, and methodologies used to train computers to "think" more like humans by simply supplying the computer with data and allowing the computer to determine the rules that govern the data. Computers will use their own processes to deduce patterns within the data. They typically uncover rules and patterns more optimized and discrete than what could be produced by humans alone. The more unbiased data a computer receives, the more it is able to course correct and further refine and improve its performance.

In short, Artificial Intelligence is a tool to easily model data developed to help humans make decisions and find patterns within complex interactions of multiple variables. Machine Learning, the application of Artificial Intelligence, gives computer systems the ability to "learn" with data, without being explicitly programed.

1-2. INDUSTRIES UTILIZING MACHINE LEARNING

Today, with many industries facing increased competition, cost pressures, market volatility and shifting customer expectations, corporations have begun leveraging internal data assets as a key driver of differentiation and innovation. Artificial Intelligence and Machine Learning are currently being used to cut costs, streamline operations, and provide better experiences and services both within an organization and for customers.

AIRLINES

In the airline industry, Artificial Intelligence and Machine Learning are currently being utilized in multiple capacities such as predicting delays, airline ticketing, supply chain/inventory improvements, and anticipating parts repair and replacements. Much of the penetration of AI into the airline industry began in 2015 with the creation of the Aerospace Data Analytics Lab, a partnership with Boeing and Carnegie Mellon University.

Boeing was generating large amounts of data. The goal: turn this data into actionable insights to create maintenance schedules based on aircraft flight history and machinery performance. No longer would Boeing have to rely on potentially inaccurate historical maintenance patterns based on other airplanes.

AUTOMOTIVE

The current autonomous vehicle race first captured the public eye when DARPA (U.S. Defense Advanced Research Projects Administration) created a $1 million challenge for the development of an autonomous vehicle capable of navigating 150 miles of desert roadway. The hope behind investing in this technology was to deploy autonomous technology in the US military by 2015. While those expectations did not quite come to pass, it did jumpstart much of the R&D that has led to the current advances in autonomous car technology. Machine Learning has long been an important component within autonomous vehicle research. As the processing power and GPUs (graphic processing units) became more powerful in recent years, the ability to intake and respond to thousands of disparate inputs and scenarios on the road became much easier to implement. Now, autonomous vehicle R&D is table stakes within the auto industry with companies like Ford and Mercedes carving out dedicated units within their organizations focused on this technology.

For non-autonomous cars, Artificial Intelligence can be used to identify causes of mechanical failure and determine if it came from an environmental or road infrastructure issue or even if it was related to a region of a country or a particular time of year. Related to this, machine learning can aid in predictive maintenance of parts and machinery, pinpointing future failures before they happen.

MEDICINE

The total healthcare sector in the United States is a $3 trillion market and projected to rise to $4 trillion by 2020. Coupled with the recent advances in supercomputing, the digitization of medical health records, and the exponential output of medical data, outpacing the growth of all other data in the total "digital universe," healthcare is quickly becoming prime ground for Artificial Intelligence and Machine Learning applications.

To bring a drug to market is roughly a 10-year process costing upwards of $1B. Much of this cost comes from drug discovery, R&D, manufacturing, and multimillion dollar clinical trials. However, many promising innovations driven by Artificial Intelligence and Machine Learning are entering this market. For example, a team of academics at Stanford University led by Dr. Sebastian Thrun, former Google VP and founder of the Google Self Driving Car program, was able to feed thousands of melanoma images into a computer vision algorithm that was subsequently able to identify benign vs. malignant lesions with the same accuracy as a medical doctor.2

Machine Learning can be applied to a number of disparate applications in healthcare including intelligent drug discovery, accurately identifying genetic anomalies, and determining the most eligible patients for clinical trials based on their condition and health history—all effectively shortening and reducing costs in the drug discovery process.

2. BASIC PRINCIPLES OF MACHINE LEARNING

In Machine Learning, data is processed in either a supervised or unsupervised manner. That is, either the computer is given several examples of patterns to look for in order to guide the analytical process or the computer is tasked with determining rules and guidelines from scratch solely based on the data set it receives.

Artificial Intelligence and Machine Learning are modeling tools that leverage critical components such as advanced computing power, computational heuristics, and mathematical intuition to analyze non-linear and oftentimes dynamic data sets. In traditional data analysis when predicting a result from an input, one of the foremost methods, linear regression analysis would often be used—that is creating a 'line of best fit.'

Given the rising amount of data that is created and available today, oftentimes detecting patterns and correlations within the data is not as easy as simply finding a single line of best fit. Sometimes principal components can only be detected by testing many different conditions and several combinations of disparate variables. These are problems that go beyond what traditional data analysis can provide. However, with enough of the right data and computational power, Machine Learning can uncover and reveal many of these complex patterns for our use.

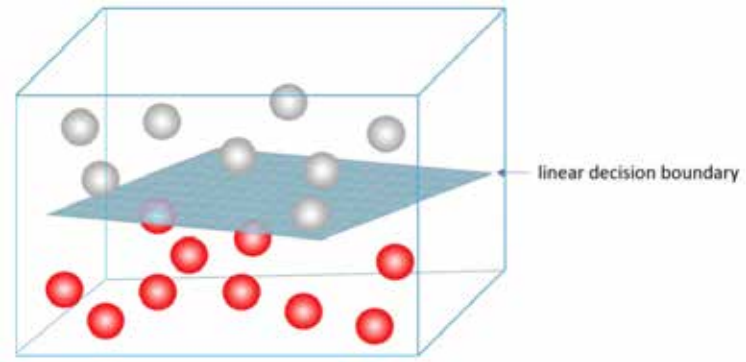
There are a number of different Machine Learning techniques. The following six are the most common:

1. Logistic Regression
2. Support Vector Machines (SVM)
3. Random Forests
4. Artificial Neural Network
5. Naive Bayes
6. K Nearest Neighbor (KNN)

Each Machine Learning technique is assessed for accuracy by running a testing data set called an AUROC.

2-1. LOGISTIC REGRESSION

Logistic regression is a commonly used classification model. When given various objects, it will determine key differences between them in order to predict the categories they belong to. For example, is it a red ball or a grey ball? Or, will a patient with a particular genetic signature respond to a certain drug? Will a student who has studied a certain number of hours pass a particular exam? Unlike linear regression models, where the output is a simple yes or no, with logistic regression the output is a probability.

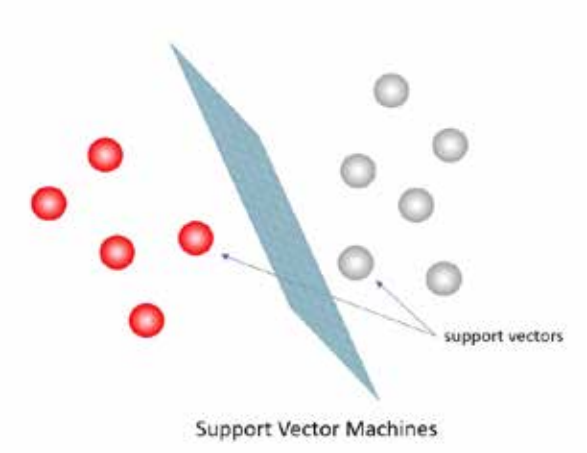


Logistic Regression

2-2. SUPPORT VECTOR MACHINES (SVM)

Support Vector Machines utilizes a similar approach to logistic regression but is known to employ a more sensitive and accurate approach when determining linear decision boundaries, or hyperplanes. Thus they are useful when categorizing objects or values that are more difficult to separate. It achieves this by utilizing what is known as support vectors, or the members in various categories that are closest to the hyperplane or decision boundary.

These models will find the hyperplane demonstrating the most separation (also known as the margin) between the hyperplane and the support vectors, indicating a cleaner and more accurate categorization. In the situation where values may not be clearly differentiable, even in a three-dimensional plane like in the figure above, the dataset will be subsequently mapped to greater and greater dimensions in order to find the dimension in which a hyperplane can accurately distinguish between two or more groups.

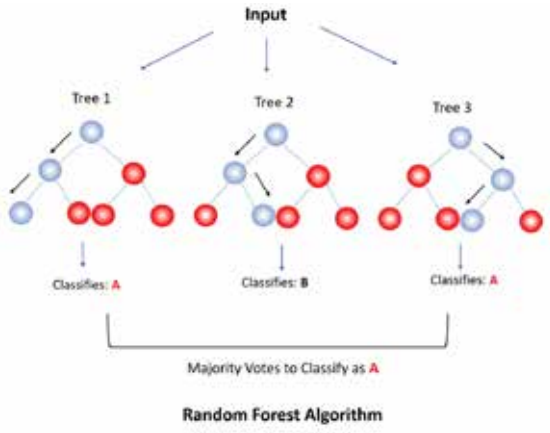


Support Vector Machines

2-3. RANDOM FORESTS

A Random Forest algorithm is a method of classification composed of a large number of individual classification or 'decision' trees. Much like a forest is comprised of a large number of individual trees, a random forest algorithm is comprised of a large number of decision trees. A decision tree is essentially a series of yes/no questions that will systematically bucket a piece of data into a specific category, and make a 'decision' about what category the input belongs to. For example, a simple decision tree may be based on a series of yes/no questions about the weather. The outcome will determine if it is suitable to step outdoors.

A random forest expands on this by incorporating additional decision trees to derive an answer. For example, the question is whether to play soccer indoors or outdoors on a particular day. One tree may ask a number of yes/no questions accounting for time of year, time of day, temperature, etc. Another tree may account for the location and slope of the proposed field. A third tree may account for weather and wind velocity. The random forest uses all three trees to derive the best answer (indoor soccer vs. outdoor). Based on the 'majority vote' a recommendation can be made.

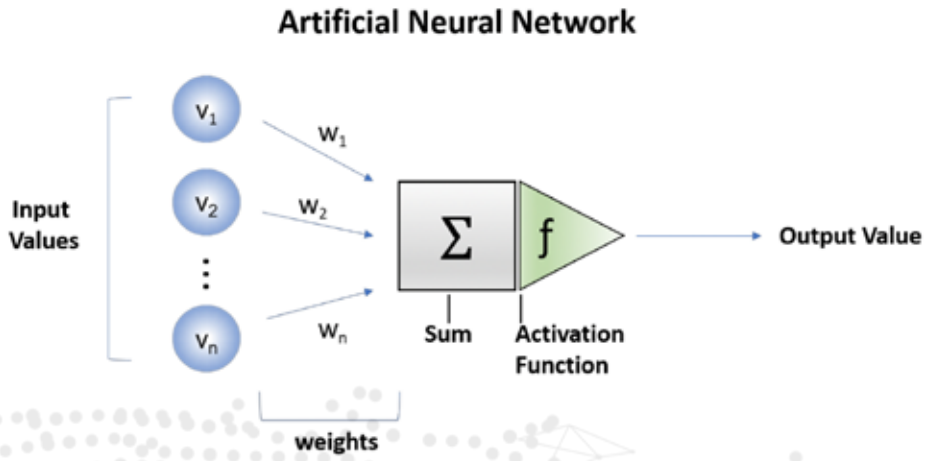


2-4. ARTIFICIAL NEURAL NETWORK (ANN)

An artificial neural network algorithm obtains its name from its similarity to the process of computation employed by brain cells (neurons). Neurons in our brains are connected to other neurons, receiving inputs from them. Depending on the strength of the collective inputs, the neuron will activate and "fire" a response.

If the strength of the collective inputs achieves a level above a certain threshold, the neuron will fire. If the signal is weak and does not reach a particular threshold, the neuron will not fire.

Similarly, in an artificial neural network, various input values outlined in the diagram above (v1, v2, etc.) are fed into the algorithm, and each is given a particular weight (w1, w2, etc.). These values and weights are then multiplied and summed in a summation step according to this equation:

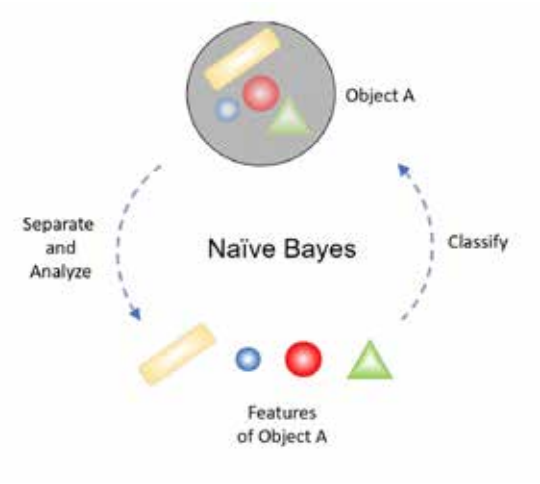


v1(w1) +(v2)(w2)+...(vn)(wn)
If the resulting value is above a particular threshold, an activation function will occur. Typically, this activation function will be a step function. That is, it will output a value of 1 if the value is above a certain threshold, and output 0 if the value falls below a certain threshold. For example, if the activation threshold for a certain neural network is set at 3, and the sum of the weighted values is 3.01, 5.7 or 294.072, then the activation function will output as 1.

However, if the sum results to a value below 3, the activation function will output as 0. The algorithm 'trains' itself by constantly adjusting the weight of a training set value depending on how much it deviates from the actual value it is trying to learn. Neural networks can be used to accomplish advanced tasks such as spam detection and even identifying fraudulent financial activity.

2-5. NAIVE BAYES

Naïve Bayes is performed by analyzing separate features of a particular object independently in order to classify the object. For example, a lemon may be considered a lemon if it is yellow, round, and at least two inches in diameter. However, these features separately by themselves would not identify a lemon.

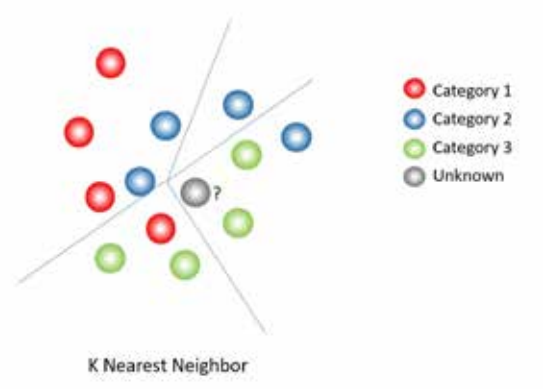


A Naïve Bayes method will correlate these features and calculate their independent probability that the object in question is indeed a lemon and not an orange.

However, sometimes features of an object are not always independent of each other which is considered one of the major shortcomings of this method and why it is termed 'naïve.' Despite this, it is one of the strongest machine learning algorithms out there, often outperforming other more sophisticated methods.

2-6. K NEAREST NEIGHBOR (KNN)

K Nearest Neighbor is an unsupervised algorithm that classifies objects based on how close that object aligns to a similar object the model previously encountered. For example, let's say you want to determine how much money a user will spend on a website. You may want to determine this before the user adds anything to their shopping cart. Understanding different demographic information (e.g. age, gender, geography) about a new user can allow one to predict, or find the nearest neighbor, of that particular user by looking at the demographic and shopping behavior of previous customers of the site.



K Nearest Neighbor

This method differs from the rote learner model in which the algorithm will only assign an outcome if the input is 100% identical to a previous historical instance. That is, unless the current object is an exact match instead of a "nearest neighbor", it will not assign that object to a specific category. For example, if an algorithm is assigning an image of a giraffe, it may not assign that image to the giraffe category unless it is pixel by pixel identical to a previous image it encountered. With K nearest neighbor, the algorithm will assign the image to a category based on how similar the image is to a previous picture of a giraffe.

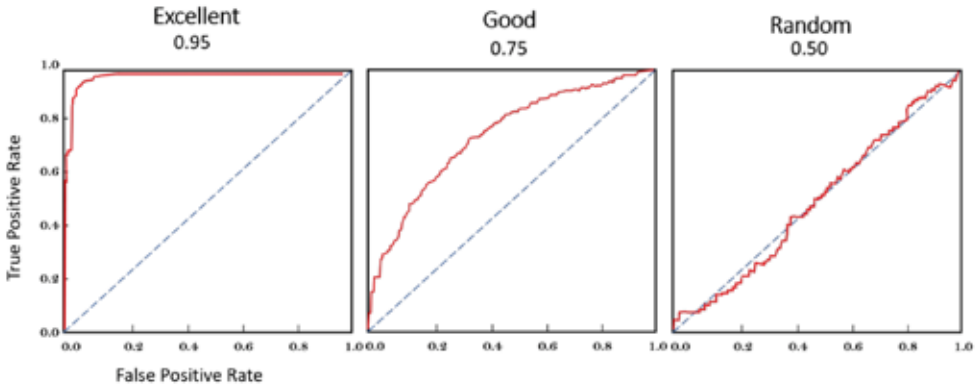
2-7. MEASURING THE ACCURACY OF A MODEL

Typically the best benchmark in determining if an algorithm will be suitable for a given problem or not is to measure the Area Under the Receiver Operator Characteristic Curve or (AUROC) sometimes simply also known as Area Under the Curve or (AUC) as shown in the figure above.

To generate an AUROC, the finished algorithm must run what is called a testing data set using unused data that was set aside. Correct predictions are called True Positives. Incorrect predictions are called False Positives.

An AUROC curve is created by plotting the instances of both True and False Positives on a graph. Starting at 0,0, each time the model predicts correctly, a point is plotted one unit up in the positive direction on the vertical axis. If the model is incorrect and selects a False Positive, a point is plotted one unit right on the horizontal axis. The area under the resulting curve is what is known as the AUROC.

The greater the number of True Positives the higher the AUROC. The higher the AUROC, the more robust and predictive the machine learning model.



Area Under Receiver Operating Characteristic Curve (AUROC)

3-1. TOOLS CURRENTLY USED FOR CONDITION ASSESSMENT

For predictive purposes, machine learning algorithms utilize vast amounts of historical data. In the water main industry, the Likelihood of Failure Analysis (LoF), also known as Condition Assessment, provides the most valuable actionable predictions.

Many utilities presume older pipes are in worse condition than newer pipes —especially older cast-iron pipes. But in fact, these pipes can be remarkably robust despite their age. On the other hand, pipes installed in more recent decades may show considerable deterioration, and could be near failure

.

It is important to use multiple variables and predictive approaches in LoF determination.

3-1. METHODS FOR WATER MAIN CONDITION ASSESSMENT

Water main pipe condition is commonly assessed using the following methods:

INVASIVE

Invasive methods necessitate the removal of pipe material or the insertion of probes. These methods typically include coupon removal for analytical testing, video inspection, coupon insertion where a piece of material is inserted into the pipe and later removed for analytical testing, and magnetic flux leakage (MFL) tools in addition to other condition assessment based PIGs (Pipeline Inspection Gauges).

NON-INVASIVE

Non-invasive methods generally involve techniques that do not compromise the integrity of the pipe, however many of these methods do still necessitate excavation adding to the time and expense involved in performing these assessments. These include acoustic velocity approaches, remote-field electromagnetic technologies, and handheld ultrasonic testing. Some of these methods are useful in detecting changes in stress level caused by soil induced bending along the length of the pipe.

DESKTOP

Desktop or computational approaches are by far the most cost effective and least invasive. The industry has adopted a number of different approaches ranging from a simple weighted score approach on an excel spreadsheet, to Cohort Analysis, LEYP, Kanew forecasting, Weibull modeling, and various commercially available condition assessment software.

3-2. DRAWBACKS OF CURRENT CONDITION ASSESSMENT METHODS

INVASIVE

Although sampling a small piece of the existing pipe and running physical condition assessment tests may seem like a direct solution, many problems that result in failures are not uniformly distributed along the length of the pipe. Instead they are more localized. Therefore, many coupons may exhibit minimal to no material deterioration, leading to false negatives resulting in a poor understanding of the sample. In addition, certain invasive methods will not be available for all pipe material types, or diameters. For example, magnetic flux leakage techniques can only be used on metallic pipes. Moreover, the turnaround time for these methods, from set-up to test results, tend to take a significant amount of time and expense. However, there are situations in which invasive methods can be useful. For example, video inspection can determine if corrosion is occurring inside or outside of the pipe reflecting the need for the application of an interior lining where corrosion can more easily be stopped.

NON-INVASIVE

Though non-invasive methodologies may be faster and somewhat less expensive than their more invasive counterparts, these assessments may not necessarily achieve a more robust and thorough assessment of pipe condition. In fact, they can suffer from similar technical limitations in that they can only assess one point of the pipe at a time and not the pipe length as a whole. As the pipe condition can vary significantly from one point to another, the best that these methods can recommend is an average degradation assessment. This is less than ideal, for example, in the case of a pipe with one or two points of vulnerability with the remaining sections otherwise sound.

Acoustic velocity methods are particularly susceptible to this limitation, as it infers stiffness of the pipe material based on detecting the speed of sound from one point to another, measuring average thickness of the pipe between the two sensor points. Thus, this method could potentially categorize a pipe with localized vulnerability as sound rather than marking it as one that will need to be quickly replaced. In addition, much like some invasive methods, non-invasive methods can also be limited to certain pipe material types and diameters, leaving organizations unable to fully assess the full extent of their total pipes.

DESKTOP

Though desktop methods may seem to be the most straightforward approach, many of these methods are based on arbitrary assumptions and weights, i.e. older pipes are more in need of replacement than newer pipes. As mentioned earlier, this is not always the case. More advanced statistical modeling may help decipher differences between variables. However, many approaches may not have the ability to consider the importance of some adjacent details such as proximity to light rails or the contribution of elevation or pipe material, factors that could impact accuracy. A more robust approach would be a large-scale comparison of these various factors to generate a more refined and accurate prediction based on the disparate interactions between component variables.

4. USING MACHINE LEARNING FOR WATER MAIN CONDITION ASSESSMENT

Due to the large amount of historical and geospatial data needed to run machine learning algorithms, water main condition assessment contains all the necessary components of an ideal application for machine learning in water utilities. The components include:

Historical data –
installation year, pipe material, and break history

Categorical data –
pressure class, geographical location, elevation and pipe diameter

Contingent data –
proximity to rail systems and soil composition

The volume of data provides a unique opportunity for water utilities. Analyzing this data consistently can uncover trends, gain insight on pipeline health, and offer data-driven assessments. Coupling likelihood of failure with consequence of failure analysis could then accurately pinpoint areas most in need of replacement.

4-1. DATA ACQUISITION, ASSESSMENT AND CLEANING

For any machine learning process, roughly 60-80% of the work is data acquisition, assessment and cleaning, also known as pre-processing or data wrangling. The remaining percentage is the machine learning itself.

1. **Data Assessment and Cleaning** (collecting and organizing the data). Data is easy to analyze and visualize. This is the heavy lifting.

2. **Machine Learning Analysis.** A dynamic solution tailored to the specific Utility resulting in accurate predictions pinpointing the location and likelihood of water main failures.

In recent years, more and more processes and paper documents have migrated to digital formats. However, a large amount of utility data and records remain in a paper and ink format as work orders and potentially misplaced engineering field notebooks. Even digital formats, such as Excel files may contain inaccurate or missing data, which can significantly affect the quality of results when using more advanced analytical processes. The good news is much of this missing data can be imputed, a process in which data can be assumed based on similarities to other pipes and historical data. For example, if a pipe diameter has the value of 0", based on the year the pipe was installed, material, and diameters of nearby pipes, a more accurate value can be substituted.

In addition to this, many of the variables present within these datasets will also need to be standardized in order for the algorithm to "learn" and predict results based on data culled from several different utilities, e.g. pipe material, pipe length, and pressure data. This essentially produces a network effect in which the accumulating data creates more accurate results for all utilities involved. Since most utilities use different nomenclatures and derivatives of similar variables, these too will need to be standardized in such a way that the computer can more accurately compare and compute similar values between utilities and construct the model beneficial to all.

Additionally, these methods also utilize critical geographic and other accessory data such as elevation, soil composition, proximity to shore and rail lines, and even water composition within the pipes – factors that would be near to impossible to include in other desktop condition assessment methods. All this is used to answer a simple question. When you have two pipes with similar histories that are exposed to different external factors, how do you identify which pipe will be more likely to fail (LoF) first?

This information enables better planning, resource allocation and preventive maintenance:

Near-term –
Likely to Fail (LOF) results allow you to defer unnecessary pipe replacement

Mid-term –
Accurately target the most vulnerable water mains in order to plan appropriately

Long-term –
Consistently lower annual break rates as overall system reliability improves

4-2. MACHINE LEARNING FOR LIKELIHOOD OF FAILURE

Once the data is assessed, cleaned, and imputed where needed, it is ready to be fed into a machine learning algorithm where it is subsequently 'trained' to learn the patterns that predict breakage events. To test the robustness of the resulting model, a process called time shifting is utilized, wherein a portion of the historical breakage data that was saved and not used during the training phase, is now used for this testing phase. Usually this process focuses on the most recent breaks. To test the accuracy of the results, a Receiver Operator Characteristic curve is plotted and the AUROC is measured (see section 2-7). The algorithm can subsequently be re-run utilizing different sets of variables until a robust AUROC is achieved. Thus with this method, machine learning can reveal insights with relatively more confidence and precision than other condition assessment methods alone.

4-3. HOW MACHINE LEARNING PREDICTION RESULTS STRENGTHEN OVER TIME

As mentioned in section 4-1, the more data a model contains, the more robust the model. As utilities collect new data over time, recording new activity, data is continually fed into a machine learning model. This subsequently enhances the model by either strengthening previously learned rules around break predictions or from encountering additional circumstances around which new rules can be built.

4-4. UTILITIES WITH MISSING BREAK OR ASSET DATA

As an ever-increasing amount of data strengthens the predictive power of a machine learning algorithm, benefiting utilities with large amounts of historical breakage and asset information, machine learning can also benefit utilities with limited asset or breakage data as machine learning data can "fill in the gaps." If a utility has little breakage data, future breaks can be informed by patterns found for other similar materials, install years, soil compositions, etc. If a utility has little asset data, a similar process can be applied by simply looking at ancillary geospatial data to impute the probability of pipe breakage events. Thus, because machine learning utilizes many streams of data in order to perform certain predictions, it begins to learn patterns that can inform situations where many of the usual data points may not be available.

4-5. HOW MACHINE LEARNING RESULTS ARE USED FOR PROCESS OPTIMIZATION

No business driven machine learning application is effective unless it is used to inform key business processes. For Machine Learning enabled LoF assessment, this data can be used to determine pipes and areas or hotspots most in need of proactive replacement. However, LoF is only part of all that needs to be considered when planning a pipe replacement job. For example, the location of the work, traffic disruption considerations and potential work collaborations with other city maintenance projects may also need to be taken into account.

Machine learning speeds up and provides greater accuracy in determining future breakage events, allowing organizations to generate better replacement plans by:

- Optimizing jobs already planned
- More easily charting future plans
- And anticipating future crew and budget resources and allocations

4-6. THE FINANCIAL BENEFITS OF MACHINE LEARNING

The financial benefits of Machine Learning models are significant. Because of its accuracy and ability to predict future breakage events, these methods can assist municipal organizations in three ways:

- The deferment of current replacement jobs

- More accuracy in targeting at risk pipes

- As machine learning goes on, break rates will decrease over time

As the algorithm can assess pipes most likely to break based on variables that go beyond simply pipe age and weighted scores, pipes that are calculated to have decades left until end of useful life can be deferred for replacements and the resources re-routed to pipes that are most likely to fail. Due to the proactive work in identifying and replacing the most poorly performing pipes, breaks rates will begin to go down as the more robust pipes will be the ones that remain.

5. CONCLUSION

Machine learning is a major trend poised to make a significant impact in a number of major industries, from airlines to healthcare, and autonomous vehicles. In addition to driving performance optimization, Machine Learning improves business processes and planning.

In the water utilities industry, due to the amount of data and variables involved, water main condition assessment is an ideal use case for this technology. For many water municipalities, water main condition assessment (LoF) is a low risk use case. The opportunity to save millions of dollars by avoiding the repair or replacement of perfectly good pipes is at last a reality.

6. COMMONLY ASKED QUESTIONS

HOW ACCURATE IS MACHINE LEARNING FOR LOF PREDICTION?

The major advantage of Machine Learning is that it is able to utilize vast amounts of disparate data, analyze it, and unlike other desktop condition assessment tools, generate predictions. An AUROC curve is the main benchmark for assessing a model's predictive power. As more data is fed into the algorithm, the more robust the algorithm becomes.

CAN PREDICTION RESULTS IMPROVE OVER TIME FOR THE SAME UTILITY?

Yes. Machine learning algorithms become more robust as more training data is available. As Utilities generate new data (i.e. pipe breaks and installation data, see section 4-3), algorithms will naturally become stronger as time goes on. As data quality improves over time, and as Utilities standardize and establish processes for more accurate record keeping, the Machine Learning model will rely on imputed data. All this taken into account will result in more accuracy.

WILL MACHINE LEARNING WORK FOR UTILITIES WITH MISSING BREAK OR ASSET DATA?

When utilities are missing key data points in their pipe or break history, that data can be imputed by utilizing information based on similar pipes, breaks, etc. from within the same utility. In the situation where a water utility may not have information regarding break history, or even pipes, machine learning data can inform potential breakage events based on data from other utilities (see section 4-4).

7. CONTACT

www.fracta.ai

1870 Broadway, Suite 200
Redwood City, CA 94063

408-901-8813

contact@fracta.ai

8. REFERENCES

1. "Worldwide Cognitive Systems and Artificial Intelligence Revenues Forecast to Surge Past $47 Billion in 2020, According to New IDC Spending Guide" www.idc.com/getdoc.jsp?containerId=prUS41878616

2. "Deep learning algorithm does as well as dermatologists in identifying skin cancer" | news.stanford.edu/ 2017/01/25/artificial-intelligence-used-identify-skin-cancer/