Practical Application of Machine Learning to Water Pipe Failure Prediction

by

Kevin Laven

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Systems Design Engineering

Waterloo, Ontario, Canada, 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Member:	Hossein Bonakdari Associate Prof., Dept. Civil Engineering University of Ottawa
Internal-External Member:	Sriram Narasimhan, Adjunct Prof., Dept. Civil Engineering University of Waterloo
Internal Members:	Chrystopher L. Nehaniv, Prof., Dept. Systems Design Engineering, University of Waterloo
	Siby Samuel, Asst. Prof., Dept. Systems Design Engineering, University of Waterloo
Supervisor(s):	Kumaraswamy Ponnambalam, Prof., Dept. Systems Design Engineering, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

As water networks age, many utilities are faced with rising water main break rates and insufficient replacement funds. Machine learning is a promising tool to support efficient water pipe replacement decisions. This thesis explores the practical application of machine learning for water pipe failure prediction using a dataset of over 10 million pipe-year records from four countries. Analysis of predictive factors shows that length, age, diameter, material, and failure history are each significant. Two novel relationships with break rate are observed: with respect to diameter, an inverse linear relationship, and with respect to age a peak at around 40 years followed by a decline lasting several decades. A method is presented for predicting both probability of failure and the expected number of failures in a given pipe and time period. By inferring units, encoding categorical features, and normalizing for different utility practices, it is proposed that a single model can generalize across utilities, geographies, and time periods without any utility-specific data cleansing. The model is trained and tested on a leave-one-utility-out basis, with training data from time periods strictly prior to test data. The resulting Area Under the Curve for the Receiver Operating Characteristic of over 0.85 and Cumulate Lift at 10% of over 5.0 demonstrate the practical applicability of the model, matching the performance of models trained and tested on each utility's own data. Within this model, a method of cross-encoding categorical features with numerical features is introduced to enable integration of data sets from diverse contributors. The applicability of these performance metrics and model outputs to common utility water main replacement decision making processes is also shown.

Acknowledgments

Completion of this thesis would not have been possible without the extensive support of my family, friends, and academic and professional colleagues.

I would like to thank my supervisor, Professor Kumaraswamy (Ponnu) Ponnambalam, for encouraging me to undertake this venture in the first instance and for the continued support and encouragement needed to complete it.

To my wife, I am deeply grateful to you for the many ways you supported me through this effort: for encouraging me to undertake the challenge, for the time you created for me to work through it, for your unwavering support and encouragement in the face of challenges, and for your acceptance of the many hours I spent away and focused on it.

My deep thanks goes out to the six water utilities who contributed the data to make this research possible. This extends both to the organizations and to the individuals therein who helped gather and arrange for sharing of the data.

Thanks also to my friends who supported me and talked me through the challenges posed by both the thesis work and the rest of the degree requirements.

Finally, I'm grateful to the many professional colleagues who taught me much of what I needed to know to complete this work and to those who helped keep enough things off my plate at work to be able to complete this.

Dedication

To my wife Marina, who made this possible.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AC: Asbestos Cement

ANN: Artificial Neural Network

AUC: Area Under the Curve, generally of the Receiver Operating Characteristic

AWWA: American Water Works Association

CI: Cast Iron

CIP: Capital Improvement Plan

CNN: Convolutional Neural Network

CoF: Consequence of Failure

CONC: Concrete

CU: Copper

DI: Ductile Iron

DN: Nominal Diameter

ELL: Economic Leakage Level

ENoF: Expected Number of Failures

EPR: Evolutionary Polynomial Regression

GBT: Gradient Boosted Tree

GIS: Geographic Information System

GRU: Gated Recurrent Unit

HDPE: High Density Polyethylene

ILI: Infrastructure Leakage Index

LOGO: Leave One Group Out

LSTM: Long Short Term Memory

MAE: Mean Absolute Error

MLP: Multilevel Perceptron

MSE: Mean Squared Error

MTTF: Mean Time To Failure

NRW: Non-Revenue Water

PCCP: Prestressed Concrete Cylinder Pipe

PE: Polyethylene

PoF: Probability of Failure

Pr: Probability

PVC: Polyvinyl Chloride

RMSE: Root Mean Squared Error (also RMS error)

RNN: Recurrent Neural Network

ROC: Receiver Operating Characteristic

RoF: Risk of Failure

SQL: Structured Query Language

ST: Steel

SVM: Support Vector Machine

XGBoost: Extreme Gradient Boosting

Chapter 1
Introduction

The distribution of potable drinking water is a cornerstone of modern sanitation. This is accomplished by using networks of pressurized pipes running from water treatment plants to individual customers. These water pipes can fail by leaking, breaking, or bursting. When this occurs, this results in the loss of treated water, damage to the surrounding area, and disruption of water supply services. Such failures are challenging to predict, as water mains are generally buried, making physical inspections difficult. This study examines the application of machine learning as a method of predicting water pipe failures. Its focus is on practical considerations in the use of such a model as a decision support tool.

1.1 Water Distribution Networks and Pipes

The main components of a water distribution system are a raw water source (usually a lake, river, or aquifer), a treatment plant, one or more pumps, and a water distribution network which brings the treated water to customers. As shown in Figure 1, a water distribution network is divided into logical segments of pipe, which typically run underground along a street and span from junction to junction. A pipe segment is often composed of many sticks of pipe, connected together by joints.

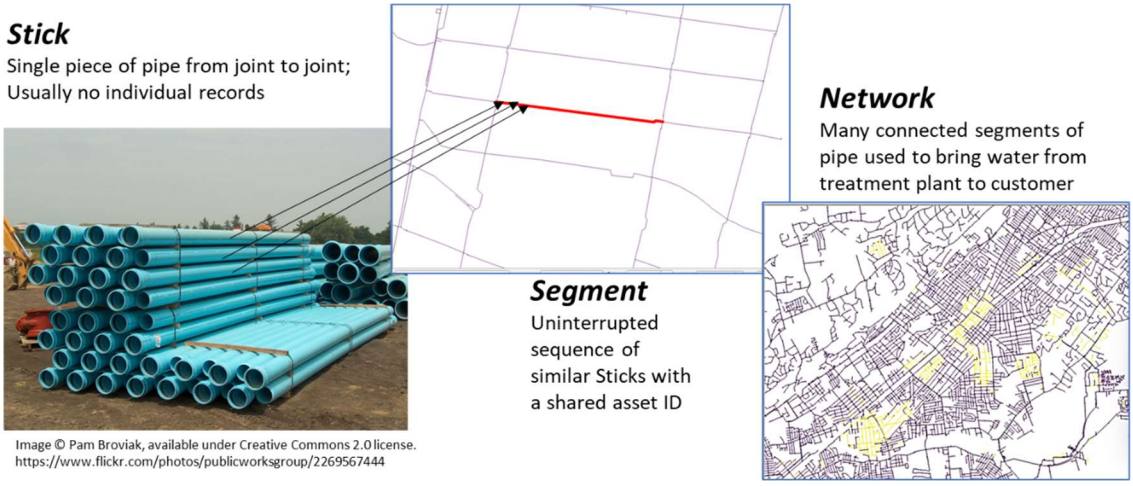


Figure 1: Illustration of a water network, pipe segments, and pipe sticks.

A range of materials have been used for pipes, such as Cast Iron, Steel, Concreate, and Polyvinylchloride (PVC). Each entire segment of pipe is generally installed at the same time, and made of a single material. Pipe segments are usually the finest grained records kept by a utility.

1.2 Water Pipe Failures and Their Prediction

A water pipe failure occurs when the pipe's hydraulic integrity is compromised. Failures encompass a wide range of severities, from tiny leaks that can only be detected via dedicated sensors, through to catastrophic bursts releasing millions of liters of water per hour. Depending on the severity of the failure, the failed pipe stick will either be repaired or replaced before the pipe segment is returned to service. A single pipe segment may fail and be repaired multiple times before being replaced.

Such failures are difficult to predict. Water pipes generally run underground, making inspections difficult and costly. Few (if any) sensors are generally present in the networks to collect operational data. This often leaves only the limited records kept during installation or during repair of past failures available for use in supporting water network management decisions.

Water networks are managed by utilities in their capital improvement programs. These programs aim to maintain the network in a state of good repair by replacing degraded or high-risk pipes. The state of repair is measured by key performance indicators such as the water loss percentage, number of breaks per 100 km per year, and the annual cost of repairing water main breaks.

Efficient management of a capital improvement program involves identifying which pipes are more likely to fail than others. This is often considered an engineering discipline. Various techniques are employed to this purpose, such as engineering judgement informed by case studies and surveys, targeted inspections and monitoring, survival analysis, failure risk assessment, financial decision optimization, and more recently the application of machine learning.

1.3 Objective of this Research

The objective of this research is to develop a practical method for the application of machine learning to pipe failure prediction. Numerous studies have been conducted in recent years into the application of machine learning for pipe failure prediction. These have either been case studies with limited generalizability, or tests of the potential for application of a particular analytical technique to the problem. None have yet attempted to meet the required criteria to be of practical use, namely:

- It must use only data to which typical utilities have access.
    - This means that any data not generally available to utilities, such as real-time sensor data and environmental data, must not be used.

- It must be applicable using only the skills available to most utilities.
- It must be demonstrated to extrapolate forward in time, and to new utilities.
- Its performance must be measured in a way that aligns with how the results will be used, demonstrating measurable value over existing processes.
- The outputs must be usable without major changes to utility engineering practices and decision-making processes.

This study aims to demonstrate that machine learning can indeed be practically applied to the prediction of pipe failures. It addresses the question of whether a single model can be used to predict pipe failures for a wide range of utilities that did not contribute data to training the model. It aims to do this without requiring utilities to manually cleanse their data and standardize to the same language, units of measure, technical jargon, and record keeping practices. It describes the process of assembling a large and diverse data set, with contributions from six large utilities in three continents, developing a new three-layer model for predicting water pipe failure risk, and then testing its ability to predict failures in new utilities which did not contribute to the training data.

1.4 Contributions of this Research

This research provides novel contributions in three areas. First are observations from the pipe failure dataset itself, being the largest and most diverse reported on to date. Second are the demonstrations of how the appropriate selection of target variables and test metrics for the machine learning problem can allow the results to be used in various engineering and utility decision practices. Third are the novel elements of the three-layer model presented.

A major element of this research project was assembling a dataset containing over 10,000,000 pipe-years of failure history. This dataset includes records from six utilities, in four countries spread across three continents. Past datasets have included data from only one country, and (with few exceptions) usually only a single utility. The size and diversity of the dataset from this project has allowed clear trends to emerge in the relationship between pipe failure rates and certain predictive variables. Some relationships, such as with pipe length and pipe material, confirmed the results obtained in smaller studies. Two novel findings were also made. First was a strong inverse linear relationship between pipe diameter and failure rate. Second was the observation of a patter of peaks and subsequent declines in the failure rate as age increases.

In reviewing the literature on the application of machine learning to pipe failure prediction, a wide range was found in the selection of target variables and performance metrics. This lack of consistency makes comparison of performance between algorithms difficult. This study observes that selection of the calibrated probability of failure in a given time interval as the target variable allows the results of a machine learning model to be directly used in a wide range of utility engineering and decision-making processes. Further, it is shown that two standard machine learning performance metrics are directly related to the impact which the results would have if used, namely the area under the curve of the receiver operating characteristic, and cumulative lift at 10%.

A three-layer model is presented as a general machine learning model for pipe failure prediction. The first layer automates the data preprocessing by inferring units of measure for numerical features and encoding categorical features. This allows data from new utilities to be used in the model without the need for manual data cleansing. The second layer is a machine learning classification model for predicting whether or not a pipe will fail in a given five-year time period. These predictions, however, are discarded, and only the relative likelihood of failure estimates used by the model are carried forward. The third layer calibrates these results to match actual failure probabilities, thereby normalizing for each utility's record keeping processes. This layer also calibrates for the expectation value of the number of failures, accounting for the same pipe failing multiple times within a time period. This allows for aggregation of the expected total number of failures within a cohort of pipes simply by adding the expectation values for all pipes in the cohort.

This three-layer model was then used to demonstrate that a machine learning pipe failure model can indeed extrapolate forward in time and to new utilities. The training and test sets were separated in time, with the training set falling prior to the test set. Model training was performed on the basis of "leave one utility out" cross validation. For each utility, the model was trained on the past data from all other utilities, and then tested on the future data from that utility. This simulates a new utility using a pretrained model built with data from other utilities. With the preprocessing later applied, the model outperformed other models which were trained and tested on each utility's data in isolation.

1.5 Organization of this Thesis

Chapter 2 reviews the literature relevant to the use of machine learning for pipe failure prediction. This review spans three broad areas. First, the construction and management of water networks and water pipes are described. This is provided both to provide a clear definition of the problem of water

pipe failure prediction, and to provide context of the management decisions which the results of such a prediction algorithm would be used to support. Second, the established engineering methods used to predict water pipe failures are reviewed. This is provided both to provide the historical context leading up to the application of machine learning methods, and also to demonstrate that, with appropriate design choices, a machine learning model can be used within these established engineering methods. Third, previous studies in the application of machine learning to pipe failure prediction are reviewed, which are directly relevant to this study.

Chapter 3 presents the methodology applied in this study. It describes the process for assembling the dataset, the exploratory data analysis, and details the novel three-layer model for pipe failure prediction.

Chapter 4 presents the results of exploratory data analysis and modeling on this study's dataset. This includes analysis of the impact of individual variables on pipe failure rates, and also the results of exploratory modeling conducted on this data.

Chapter 5 presents the results obtained by applying the generalized machine learning model for pipe failure prediction.

Chapter 6 presents a discussion of the results obtained, including results from the exploratory analysis of the dataset and from the new model proposed, as well as the potential relevance and impact of these results.

Chapter 7 presents the conclusions of the study, along with limitations of the model and directions for future research.

Chapter 2
Literature Review

2.1 Water Pipe Network Construction and Management

This section provides a background on the construction of water networks, and the management decisions made by water utilities related to water main replacement.

2.1.1 Pipe Construction and Failures in Water Networks

According to a 2020 American Society of Civil Engineers report, the United States projects a $2.2 trillion dollar water infrastructure spending gap between 2019 and 2039, spending roughly $1 trillion out of a $3.2 trillion need (DiLoreto et al., 2020). This gap arises because much of the water infrastructure is out of sight and out of mind: buried underground. The fact that a pipeline has degraded becomes apparent only once the main bursts. While water main breaks are rising (Folkman, 2018), this effect is gradual and does not thrust the problem into the public eye.

With insufficient capital funds to replace all infrastructure that surpasses its design life, utilities must be selective in replacing or rehabilitating only those water mains that have, in fact, degraded. Yet the fact that water mains are buried underground also prevents utilities from directly assessing their condition. Instead, utilities must rely on limited data to infer the conditions of water mains and project these into the future.

Addressing this contradiction is the essential challenge posed by pipeline diagnostics. With limited available data, determining which pipes need to be rehabilitated or replaced to maintain quality of service is usually measured by the water main break rate. Available data is usually limited to pipe demographics (age, material, etc.), maintenance records, and minimal (if any) sensor or testing data.

Historically, this diagnostic problem has been addressed via human judgment. Each utility creates its own formula to assign condition ratings based to pipes. These formulae generally consist of subjective rules, often as simple as age. More complex rules incorporate other factors, such as material and diameter, in some cases based on an understanding of the failure mechanisms of the pipes in question.

Studies have investigated the impact on failure rates of individual factors in individual utilities. These valuable contributions enable more data-driven rules for condition ratings. These condition ratings can be combined with the consequence of failure ratings to yield net risk ratings for each pipe.

Yet these approaches remain qualitative. An alternative approach, rarely practiced due to a lack of readily available methods, is to make quantitative estimates of the probability (or expectation value) and consequence of future failures and to use these to calculate the expectation value of future failures.

The problem to be investigated in this project is providing quantitative methods for estimating the probability or expectation value of future failures using machine learning methods.

Machine learning techniques offer an avenue for making quantitative estimates of failure probability that consider all the available data on each pipe. Few studies have been attempted in this domain. These have generally been limited to data from a single utility and have predominantly relied on basic machine learning techniques that do not capture the relationship between pipes nor the time relationship between events on pipes.

Advances in machine learning over the past decade have dramatically improved performance on many problems. A review of the recent use of these techniques in the related problem of medical diagnostics highlights approaches which may be effective in pipeline diagnostics.

Readers should note that the author has worked in the field of pipeline diagnostics and failure management from 1999 through 2018. Portions of this thesis related to the management of water pipelines and their failures incorporate generally accepted industry knowledge and perspectives developed from 20 years of work and study in this area.

2.1.1.1 Water Networks and Pipes

2.1.1.1.1 Pipes in Water Networks

A water network consists of a set of interconnected pipes that bring water from sources to consumers. To be considered part of the same network, the pipes must be hydraulically connected to each other; that is to say, there must be a path through which water could flow from any given pipe to any other given pipe within the same network. Networks can be broken down into four components: transmission mains, trunk mains, distribution mains, and customer connections.

Transmission mains transport water in bulk from one location to another. They generally run from raw water sources to treatment plants and then from treatment plants to a large group of consumers (typically a town or city but sometimes a single large consumer such as a power plant). They tend to be large diameter (typically 500mm to 3000mm, sometimes even larger) and run for long distances without junctions or isolation valves.

Trunk mains are larger diameter water mains that move water about within an urban center. Some water networks are built in a "trunk-and-branch" arrangement, with larger trunk mains bringing water to zones or areas in the network and then smaller mains distributing water within that zone or area. Other networks employ a grid or loop system to provide redundancy. The term *trunk mains* is still sometimes used to describe the larger diameter mains within a grid or loop system. Trunk mains are generally of medium to large diameter (typically 200mm to 1000mm). While they tend to run moderately long distances between junctions with other trunk mains, junctions with smaller diameter distribution mains or isolation valves can be more frequent. A few large customers (hotels, hospitals, factories, etc.) may receive a connection directly from a trunk main.

Distribution mains are the small diameter water mains that bring water supply into close proximity for most customers. They are generally small diameter (50mm to 150mm) and tend to run short distances between junctions and isolation valves. Distribution mains almost always follow streets, laid either underneath the street or in a right of way along the street. Most customer connections are attached to a distribution main.

Customer connections are the individual service lines connecting a single customer to the distribution network. They often consist of three components: a water meter with an isolation valve (called a curb stop) connected to it, a pipe leading from the meter to the customer building (the customer-side pipe), and a pipe connecting the meter to the distribution main. It is common practice (but not universal) for the customer to own the pipe on their side of the meter and the utility to own the pipe connecting the meter to the distribution main. Ownership of the meter itself and the isolation valve varies. Each full set of components (utility-side pipe, curb stop, meter, and customer-side pipe) is generally referred to as a single customer connection or customer supply pipe.

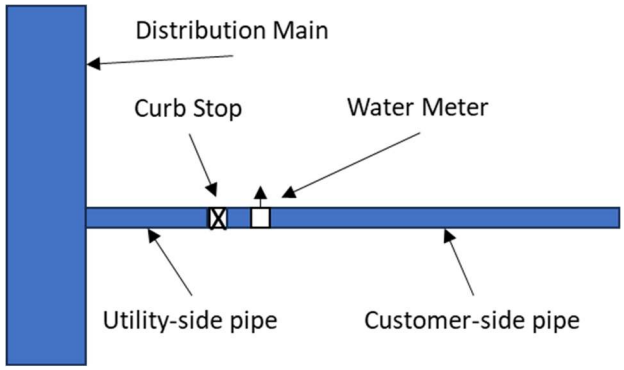


Figure 2: Components of a customer connection.

2.1.1.1.2 Defining a Pipe

The notion of pipes is simple and intuitive: cylindrical vessels used to transport fluid from one place to another. It is clear that a network consists of many different pipes. What is less clear is what constitutes a *single* pipe within the network, as the term *pipe* is ambiguous.

Here, we introduce three specific terms: *stick*, *segment*, and *pipeline*. Each of these represents a different level of granularity. The term "pipe" can sometimes refer to each of these terms, both industry and literature. The reason for this is that a pipe is a logical entity rather than a physical entity.

The finest reasonable level of granularity is to define a single *stick* of pipe as a member of the population. Most types of pipe are produced in discrete lengths (typically 4 to 6 meters in length). These are jointed together, frequently in a bell-and-spigot configuration, to produce pipelines. For some types of pipe, common degradation mechanisms tend to be isolated to a single stick of pipe. For example, prestressed concrete cylinder pipe (PCCP) tends to fail due to corrosion or embrittlement of the prestressing wires. This corrosion does not naturally spread from one stick of pipe to the next. As a result, assessment and rehabilitation tends to occur on a stick-by-stick basis. For such types of pipe, individual sticks are the natural "members" of the population. The challenge with using such fine-grained definitions is that associating a data point with a particular member is challenging. Pipeline design drawings are prone to errors, as the actual installation often differs from the original design. Even when correct, making the connection between physical location and the "station numbers" used to describe a location along a pipeline is often difficult. Even for PCCP pipes, when a particular stick of pipe is selected for reinforcement, it is common practice to reinforce

the sticks on either side of it as well, as the odds of misalignment between the analysis and the physical pipeline are high.

Another reasonable level of granularity is to consider a *segment* of pipe - the length of pipe between adjacent "nodes" in the pipe network - as a member of the population. These nodes often consist of intersections, isolation valves, pressure reducing valves, or other features that offer a physical separation along the pipeline. Some types of plastic piping material are produced in long spools rather than individual sticks. These types of pipe are typically installed in one continuous stretch between nodes in the pipe network. For such pipes, the lengths of pipe between nodes in a network are the natural "members" of a population. This definition is often used in the Geographic Information Systems (GIS's) used by utilities to track their underground infrastructure. This approach does, however, come with four drawbacks. The first is that there is no natural way to include the "nodes" themselves in the model. These nodes are common failure points, making this a significant drawback. The second is that there is no universal definition of what separates one segment from another. For example, some utilities consider the short length of pipe connecting a fire hydrant to a distribution main to be a pipe itself, whereas others consider it a feature on the distribution main. The third, which is partially a consequence of the second, is that the size of the population members varies widely, which can pose challenges to diagnostic models. A single "pipe" that is 800m long has far more locations, each of which could experience a leak or burst, than one that is 20m long, regardless of its condition. In regression models, this can be dealt with by normalizing for length (e.g., predict the number of leaks per km), but classification models have been prone to ignoring this factor. Finally, members of the population defined this way are not necessarily persistent over time. A long stretch of pipeline can be "split" into two by adding a new valve in the middle of it.

Differences in rules for defining a single segment can be of material significance when performing analysis related to water main replacements. While there are common conventions to this grouping (e.g., a junction separates pipes, an isolation valve separates pipes, a change in material separates pipes), there are also variations among utilities. A prime example of these differences comes from the use of repair pieces. When a pipe leaks or breaks, it is generally either repaired with a clamp or the broken section is cut out and replaced with a repair piece (a single stick of pipe, often shorter than a standard stick and usually of a different material). When a repair clamp is used, the standard

convention is that the pipe remains the same pipe. However, when a repair piece is placed in the middle of pipe, the pipe records can be updated in one of several ways:

a) No change; the pipe remains the same pipe.

b) The pipe is split into two, with two new pipes created.

c) The pipe is split into two, with one portion remaining the original pipe and one new pipe created.

d) The pipe is split into three (the upstream section, the repair piece itself, and the downstream section), with three new pipes created.

e) The pipe is split into three (the upstream section, the repair piece itself, and the downstream section), with one portion remaining the original pipe and two new pipes created.

This is a recordkeeping policy decision for each utility, and it can have significant implications for pipe break forecasting. Take the example of a 200m long pipe segment that has suffered two leaks at the 50m point and the 150m point along its length, both of which were repaired with clamps. A new break forms at the 100m point that requires a repair piece. If the utility follows policy a) above, then there is a single pipe with three breaks in its database. If the utility follows policy c) above, however, then there could be a pipe with three breaks and another with zero breaks.

Finally, we can consider a *pipeline.* This is a group of one or more *segments* that constitute a single path from a source to a destination. The pipeline is generally of a single material and of a similar diameter throughout. For example, a single pipeline may start at a water treatment plant and run to a reservoir in the center of a city. There may be several isolation valves along its length, several branch lines drawing water from the pipeline between the start and end, and it may even change in diameter if these branch lines draw significant amounts of water.

For the purposes of this study, we will use the term *pipe* to refer to a segment. This is the unit of granularity most commonly used for making pipe replacement decisions, making it appropriate for this study. The terms pipe and segment may be used interchangeably herein. The terms *stick* and *pipeline* will be used explicitly when referring to these granularities.

2.1.1.2 Pipe Materials and Their History

There are eight common pipe materials in use for water pipes: cast iron, steel, concrete (including reinforced concrete and prestressed concrete cylinder pipe), ductile iron, asbestos cement, polyvinylchloride (PVC), polyethylene (PE) including high density polyethylene (HDPE), and copper. These are described in this section, together with their history of use. Except where otherwise referenced, the content can be found in the American Water Works Association (AWWA) M77 guidebook (Ellison et al., 2018). A summary is provided in Table 1, with details provided in Appendix A.

Table 1: Summary of common pipe materials and properties.

Material	Acronym	Diameters	Usage Period	Key Note
Cast Iron	CI	Small – Med	1700 - 1960	Pit cast and later spun cast
Ductile Iron	DI	Small – Med	1960 - 2000	Cement linings now common
Steel	ST	Med – Large	1850 - present	Common for critical pipes
Asbestos Cement	AC	Small – Med	1940 - 1970	Many utilities removing
Polyvinyl Chloride	PVC	Small	1950 - present	Earlier use in Europe
Polyethylene	PE/HDPE	Small	1950 - present	Includes HDPE
Concrete	CONC	Large	1950 - present	Includes PCCP
Copper	CU	Small	1940 - present	Mainly customer service lines

2.1.1.3 Pipe Degradation and Failure

Like any mechanical structure, a pipeline will fail when the stresses placed on the pipeline exceed the strength of the pipeline. Pipeline failure management consists of assessing the current condition of the pipeline, including various measures of strength and stress, and applying a judgement as to the current (or a projection as to the future) fitness for service based on these.

Most pipelines are buried, making them difficult to observe directly. As a result, assessing the current condition of a pipeline is a diagnostic problem requiring inference from indirect measures.

Another diagnostic problem of interest is projecting a pipe's future fitness for service. Pipeline management decisions, such as selecting pipes for replacement or rehabilitation, are generally made based on projections of a pipe's future fitness for service. This fitness for service can be quantified by the pipe's failure rate (breaks per pipe per year), normalized failure rate (breaks per 100 km per year), or time to next failure. This diagnostic problem involves projecting a pipe's current conditions into the future. As noted above, the actual current condition of the pipe is rarely known. As a result,

projecting a pipe's future conditions is, in fact, a compound problem of first inferring the pipe's current condition and then projecting it to the future.

This section introduces the mechanisms of pipeline degradation and failure as well as the signs and symptoms of progression towards failure.

2.1.1.3.1 Degradation and Failure Mechanisms

The definition of a pipeline failure is not universally agreed upon in the industry. The terms "burst," "break," and "leak" are each used to describe a pipeline that is allowing fluid to escape. These terms are sometimes used interchangeably. When used to differentiate severity, it is usually the case that burst >= break >= leak. Yet the state in which a pipeline has "failed" differs depending on the situation. A pipeline transporting raw water in a water-rich region may have a hole losing hundreds of gallons per minute and be considered fit for service. Conversely, a pipeline through a water scarce region carrying water treated by desalination may be deemed to have failed if even a miniscule leak is present.

While the state at which failure occurs is contentious, the path towards this state is better understood. Pipelines are designed to withstand their expected stresses, plus some safety factor. According to a 2005 Environmental Protection Agency white paper, "Water main breaks are caused when and where the loading on the pipe exceeds the pipe strength (i.e., ability to resist loading)" (Royer, 2005). If a pipe initially operates as expected but eventually fails then something about either these stresses or the pipe's ability to withstand the stresses must have changed over time. The path to failure may thus involve increases in stress, loss of ability to withstand stress or a combination thereof (Wilson et al., 2014).

Perhaps the best understood path to failure is corrosion. In ferrous pipes, the internal or external surface will corrode with time. This corrosion reduces the thickness of the structurally sound material, which is referred to as the wall thickness or structural wall thickness. This loss of wall thickness represents a reduction in strength for the pipeline (Rajani & Makar, 2000). It can spread broadly across the entire surface (e.g., from acidic water or soil) (McFarland et al., 2012) or locally (e.g., pitting corrosion, galvanic corrosion, stray current induced corrosion) (Baird, 2011). Corrosion protection coatings are often applied to slow this process. These coatings themselves will degrade and fail over time, which can also happen either locally (via mechanical damage) or more broadly with time. Corrosion can also occur in the ferrous portions of composite material pipes, such as when

the steel reinforcing wires in prestressed concrete cylinder pipe (PCCP) corrode and break (Ge, 2016). Asbestos Cement (AC) pipe can undergo an analogous process whereby the chemical composition of the pipe wall changes over time in a manner that reduces its strength (Ghirmay, 2014). This is sometimes referred to as effective wall thickness loss. Corrosion often results in gradual failure, where a pipeline leaks before failing.

Another recognized path to failure is the application of severe, unplanned stress to the pipeline. Examples of stresses that have been studied include third party damage (Makar et al., 2001), pressure transients (Lambert, 2000), frost heave (Selvadurai & Shinde, 1993), soil shifting (Makar et al., 2001), and traffic loads (Stone et al., 2002). These stresses sometimes cause a pipeline to fail suddenly.

These two paths to failure are by no means mutually exclusive. For a pipeline to fail, it is sufficient for the stresses to exceed the strength only at one point, and in one direction. An unplanned stress that a new pipeline could withstand may cause a degraded pipeline to fail. Even when an unplanned stress is severe enough to cause a new pipeline to fail, this failure will occur at the weakest point on the pipeline.

2.1.1.3.2 Leakage: Existing Failures

Along with playing a role in defining whether a pipe has already failed, leakage can be a sign of impending failure (Chastain-Howley, 2005). The formation and growth of leaks tends to follow one of two broad paths. One path is characterized by the formation of a pinhole-sized leak, which grows slowly and steadily over time. The second path is characterized by crack-like leaks, which can form and grow in sudden episodes (Rajani & Kleiner, 2012).

A more detailed explanation of the leakage phenomenon and related inspection technologies, based on the author's 15 years of professional experience in this field, is provided as optional supplementary reading in Appendix A. That material is outside the scope of the proposed research project. For this project, it is sufficient to observe that past leakage is recognized as an indication of future leaks and breaks.

2.1.1.3.3 Pipe Condition: Indicators of Future Failures

While leakage identifies pipes that are already failing to hold fluid as designed, pipe condition describes the propensity for future failures by via early signs of degradation. While small leaks can

themselves be signs of degradation, more often condition assessment aims to detect loss of strength in the pipes prior to leakage.

The results of condition assessment projects fall into two types: condition ranking and quantitative assessment (Ellison et al., 2018).

Condition rankings (e.g., “A through F” or 1 to 10) are subjective ratings based on utility staff experience. They are generated using rules applied to available pipe records (age, material, past failures, etc.), with each utility creating their own rules from staff experience or engaging an engineering consultant to create rules for them. No standard for these rankings exists; however, the consistency of the rules used by different utilities has been demonstrated by training a single artificial neural network to predict the rankings to pipes by three different utilities (Al-Barqawi & Zayed, 2006).

Quantitative assessment involves estimating a measurable value. Values such as the remaining wall thickness, number of broken prestressing wires (Atherton et al., 2000), maximum corrosion pit depth, current burst strength or crush strength, future probability of failure, and the time to next failure are examples of quantitative assessment. The specific techniques and technologies used to assess pipe condition are described in Chapter 2.2.1.

2.1.1.4 Management of Pipe Failures

Pipe failures are generally treated as an inevitable consequence. Utilities treat pipe failures as something to be managed, rather than entirely prevented. It covers how the failures themselves are dealt with, the costs (both direct and indirect) associated with failures, and the norms and targets that exist around the world.

Management of pipe failures is done based on objectives or KPIs set by utilities and their regulators. The common objectives are water conservation, financial efficiency, and customer service.

Where water conservation is the objective, common practice is for utilities to track their pipe failure performance based on the Non-Revenue Water (NRW) percentage (Ellison et al., 2018). This is calculated as:

$$NRW = \frac{SIV - WSV}{SIV} \quad \textbf{(1)}$$

Where:

- *SIV* = System Input Volume: bulk purchases plus raw water extraction
- *WSV* = Water Sales Volume: metered water sales, plus estimated unmetered sales

While NRW percentage is the most common metric used, it is not the recommended metric. The reason for this is the presence of SIV in the denominator. An effective program by the utility to encourage its customers to conserve water will result in reductions in the SIV, which will leave the numerator as is but reduce the denominator. Consequently, the NRW percentage will increase due to the utility running a successful conservation program. Alternate KPIs encouraged by industry associations include leakage in L / customer connection / day, leakage in L / km of pipe / day, and the Infrastructure Leakage Index (ILI) (Jernigan et al., 2019).

Where the objective is financial efficiency, the concept of the Economic Leakage Level is used. This concept considers the tradeoff between the cost of efforts to reduce leakage and the cost of treating and pumping additional water that ends up leaking. An example of the Economic Leakage Level concept is the Economic Intervention Frequency. This concept considers the cost of proactive leak detection programs run at some time interval t. As t decreases, the annual cost of the intervention increases. As t decreases, the average run time of an undetected leak decreases. The Economic Intervention Frequency is the interval t that produces the lowest total average annual cost (Lambert & Lalonde, 2005).

Where customer service is the objective, a common performance metric is the number of breaks per 100 km per year. A water main break represents both a disruption to water service for the end customers relying on that water main, as well as a disruption to traffic. Common targets are to keep water main breaks below a threshold of 20 breaks / 100km / year, or 30 breaks / 100 miles / year for utilities that use imperial units (Ellison et al., 2018).

2.1.2 The Aging Pipes Decision: Replace, Reline, Rehabilitate, or Run to Failure?

As pipe networks degrade, rates of main breaks can increase to levels beyond the targets set by or for utilities. There are four essential options for handling an aging (and potentially degrading) pipe:

replace the existing pipe with a new one, reline the pipe entirely, rehabilitate selected sections of the pipe, or run the pipe until it reaches a failure state.

2.1.2.1 Pipe Replacement

Pipe replacement involves laying a new pipe in the original location of the old pipe.

Historically, pipe replacement was done via an open-cut approach. Under this approach, a trench is excavated around the location of the existing pipe. The old pipe material is removed for disposal, recycling, or to be sold as scrap metal. A new pipe is laid in the original trench and then returned to service. Open-cut methods allow new materials or different diameters to be laid. A large portion of the cost of this installation method lies in the excavation and restoration of the surface cover, particularly when the pipe ran underneath roadways. Coordination of rehabilitation activities with other underground utilities (sewers, gas lines, power lines, etc.) and/or with planned road resurfacing can reduce the cost.

Trenchless methods for pipe replacement have become more common over the past several decades (Wu et al., 2021). Horizontal directional drilling allows a new pipe to be run between two pits, avoiding the cost of excavating along its full length. This generally requires a non-jointed pipe material, such as HDPE, so that the new pipe can be pulled or pushed into place. If a smaller diameter is acceptable for the new pipe, then the pipe-in-pipe technique can be used, whereby the new pipe is drawn through the old pipe. Pipe bursting is similar to pipe-in-pipe, except that a burst head is forced through the old pipe prior to drawing through the new pipe, allowing a new pipe of the same or even slightly larger diameter to be used.

2.1.2.2 Pipe Relining

Aging pipes will sometimes have their service life extended by relining the pipe. This process generally involves cleaning the pipe first to remove any debris, scale, or corrosion product that has built up on the inside pipe wall. A liner is applied with the intention of extending the service life. Liners can generally be divided into two groups: structural and non-structural.

Non-structural liners aim to restore the hydraulic integrity of a leaking pipe, and to delay the onset of corrosion caused by contact between the water and the pipe wall. Internal cement mortar linings are intended to prevent or delay internal corrosion. Spray-on linings have been applied to cast iron and ductile iron pipes since the 1940s, and it has become common for iron pipes to have a cement

mortar lining pre-applied during manufacture, preventing the need for application (Dąbrowski & Li, 2021).

Structural liners aim to provide additional structural strength to a pipe, in addition to corrosion protection. These liners are often made of plastic or resin. They are either sprayed on or drawn through the pipe and then cured in place (Wu et al., 2021).

2.1.2.3 Pipe Rehabilitation

Pipe rehabilitation describes the process of extending the service life of a pipe without wholesale replacement or relining. This generally involves testing or monitoring the full length of the pipe, and applying preemptive repairs or reinforcement at locations deemed at high risk of failure (Zarghamee et al., 2011).

There are two common approaches to rehabilitation. One is inspection and spot reinforcement of large diameter pipes, whereby an inline inspection device is run through the pipe, and at-risk locations suffering from degradation are reinforced. The second is leak monitoring and spot repairs, whereby a monitoring system is used to detect small leaks as they form on the pipeline, which are then excavated and repaired prior to the leak growing into a larger burst.

2.1.2.4 Running Pipes to Failure

The final option is simply "running the pipe to failure." While this option is commonly selected by utilities for pipes with a low consequence of failure, it means different things to different utilities due to the absence of a universally accepted definition of the failure state for a pipe segment. This practice is discussed further in Chapter 2.2.3.5.1.

2.1.3 Replacement Decision in Organizational Context

Pipeline replacement decisions are made within the broader planning and operational context of water utilities. There are three common planning processes in use at water utilities. Each has its own cadence, its own time horizon, and its own specific decisions to be made.

2.1.3.1 Long Term Decisions: Master Plans

Master Plans describe a utility's long-term plan for providing services in a reliable and cost-effective manner. They typical look 25 to 50 years into the future (GM BluePlan Engineering, 2020). The

core question here is “how much should be budgeted for pipe replacement?” Examples of specific decisions that can be aided with a pipeline failure prediction model at this stage include:

- How much capital will we need for pipeline replacement over the coming decades?
- What design decisions can we make (i.e., pipe material & diameter) that will help to keep costs down?
- Should we focus on replacing or rehabilitating aging mains?
- What balance of capital spending (i.e., replacement) and operational costs (i.e., repairs & maintenance) offers the best long-term impact?

2.1.3.2 Medium Term Decisions: Capital Improvement Plans

Capital Improvement Plans outline specific projects to be undertaken in the upcoming period of three, five, or ten years. They are typically constrained by policies in the Master Plan, and by available capital budgets. These plans are where tactical decisions are made regarding management of specific pipelines. The core question here is “which pipes should be replaced in the coming years?” Examples of decisions that can be aided with a pipeline failure prediction model at this stage include:

- Which pipes should I prioritize for replacement over the next five years to minimize the future break rate?
- Which of the prioritized pipes should be replaced, and which should be relined?
- Which pipes offer the best return on investment for replacement?
- Which pipes are more economical to rehabilitate than to replace?

2.1.3.3 Short Term Decisions: Operational Plans

Operational planning is generally done on an annual basis as part of the budget planning cycle. Often looking forward one to three years, these plans set out the likely expenses to maintain services. Because these operational plans are made within the constraints of the pipeline network, there are a limited number of decisions available which could be informed by a model. Some are:

- How much should I budget for main break repairs in the upcoming one to two years?
- Are we better off using contractors or utility staff for main break repairs?

2.1.4 Specific Decisions to Be Addressed by This Study

This study aims to provide a practical method of using machine learning to support certain water main replacement decisions. The decisions the method should support are grouped into two categories: questions that are addressed directly by the model and questions where the model supports obtaining answers together with additional analysis.

2.1.4.1 Questions to be Answered Directly by the Model

The model outputs should provide direct answers to the following questions. These questions are primarily related to the Medium-Term time horizon (Capital Improvement Planning). They are not an exhaustive list but rather reflect a representative sample of questions which should be answerable.

- Which water mains are most likely to suffer a break in the next five years?
- How likely is it that a particular pipe will break in the next five years?
- How many breaks can be expected from a particular group of water mains in the next five years?

2.1.4.2 Questions For Which the Model Supports Answers

The models and their outputs should, when used together with additional analysis methods, provide useful support to answering the following questions. These questions are primarily related to the Long-Term time horizon (Utility Master Planning). They are not an exhaustive list but rather reflect a representative sample of questions which should be answerable.

- How many water main breaks per year should we expect in each five year period for the next 25 years or 50 years?
- Which pipe materials and diameters offer the best long-term value?
- How much do we need to budget for water main replacement to keep water main break rates below a threshold of 20 breaks per 100 km per year?
- Which groupings (cohorts) of pipe will be cost-effective to replace in the next 25 years?

2.2 Established Engineering Methods for Water Pipe Failure Prediction

This section outlines the established engineering methods for water pipe failure predictions, and how these are commonly used as supporting methods for pipe replacement decisions. The methods are discussed in the context of the specific management decisions outlined in Chapter 2.1.4.

2.2.1 Inspection and Monitoring

One approach used to predict pipe failure is applying various inspection and monitoring technologies to the pipes. While effective, these techniques require specialized equipment and expertise, making them costly and time consuming. As a result, these are often used only in a targeted manner rather than broadly across water distribution networks.

The technologies and analytical methods commonly applied vary depending on the pipe material in question. This section of the review offers a summary of the common pipe types and condition assessment methods, their applicability as summarized in Table 2, and summary descriptions below. Greater detail is available in the AWWA M77 guidebook (Ellison et al., 2018).

Table 2: Common pipe types and condition assessment methods.

	CI	DI	Steel	AC	PVC	HDPE	CONC	CU
Pit depth measurement	Yes							
Ultrasonic Testing	~							
Near Field Eddy Current		~						
Magnetic Flux Leakage			~					
Remote Field Eddy Current			Yes					
Transformer Coupling							Yes	
Wire Break Monitoring							Yes	
Acoustic Wall Thickness Testing	Yes	~		Yes				
Phenolphthalein Dye Testing				Yes				
Leak Detection		Yes	Yes			~		
Inline Video (CCTV)								

2.2.1.1 Common Condition Assessment Methods

This section introduces the various common methods of condition assessment for water pipelines. Much of this material is drawn from the personal experience of the author, based on over 15 years of experience developing, commercializing, and applying condition assessment technologies. A summary is provided in Table 3, with details provided in Appendix A.

Table 3: Summary of condition assessment methods.

Method	Applicability	Usage	Raw Results	Analyzed Results
Pit depth measurement	CI	Samples	Pit depths	Statistically extrapolated maximum pit depths
Ultrasonic testing	CI, DI	Local	Pulse times	Wall thickness at test point
Near field eddy current	CI, DI, ST	Local / inline	EM signal	High resolution wall thickness
Magnetic flux leakage	CI, DI	Local / inline	Flux at each sensor	Pit depths
Remote field eddy current	DI, ST	Inline	EM signal	Pit depths
Transformer coupling	PCCP	Inline	EM signal	Broken prestressing wires
Wire break monitoring	PCCP	External	Acoustic signal	Breaking prestressing wires
Acoustic wall thickness testing	CI, AC, DI, ST	External	Acoustic signal	Average wall thickness over interval between sensors
Phenolphthalein dye testing	AC	Samples	Cross section images	Wall thickness in cross section
Leak detection	All	External	Acoustic signal	Leak location
Inline video (CCTV)	Rare	Inline	Video signal	N/A

2.2.1.2 Application of Machine Learning to Leak Detection

While the focus of this study is predicting future breaks in water pipes, a closely related problem is identifying and locating existing breaks (i.e., leaks) that are not visible from the surface, a process known as leak detection or leak monitoring. Practical application of machine learning to this problem includes using acoustic signals from sensors either temporarily placed (leak detection) or permanently installed (leak monitoring) on the pipe, usually for this specific purpose. Another practical approach is inducing small pressure transients (also known as water hammers), measuring the propagation speed and head loss, and comparing these to the expected speed and head loss, which can be

calculated using various numerical methods (Bostan et al., 2021). Another approach involves using flows and non-transient pressures to identify the specific locations of leaks. The latter approach is effective in identifying isolated zones or areas in a network (known as District Metered Areas, or DMAs) experiencing heavy leakage. The application of this approach to locating leaks more precisely is strictly theoretical at this time. This section provides a brief review of selected studies in the literature on this topic.

The task of detecting leaks using acoustic vibration data has been shown to be effective when a Gaussian Mixture Model is applied (Cody & Narasimhan, 2020). Accuracy rates of over 70% were achieved in discriminating simulated leaks from non-leaking pipes.

One study investigated the use of deep learning for determining the location of a burst using data from pressure meters operating at 15 minute intervals (Zhou et al., 2019). This study showed promising results but relied entirely on synthetic data (simulated in a hydraulic modeling package) for both training and evaluation of the algorithms. This reliance is a substantial weakness in the study, as it is in practice simply demonstrating the ability of a neural network to approximate the inverse functions of the equations used in the simulation.

Fan and Yu applied a two-step approach of clustering to define pressure sensor placements followed by training machine learning models for detecting whether or not a leak is occurring in each zone (Fan & Yu, 2022). They concluded that with optimized sensor placement, only a small number of known and measured historical leaks were needed in the vicinity of each pressure sensor. Using data from two hypothetical water networks simulated in a hydraulic model, over 95% of leaks could theoretically be detected, and over 80% could be localized.

Cai et al. employed support vector machines to classify pipes in a simulated network as leaking or not leaking using only simulated flow and pressure measurements and were able to achieve an AUC score of 0.891 on this simulated, calibrated, noise-free network (Cai et al., 2022). The authors also applied a multi-stage approach of prioritizing areas of a network, detecting the presence of leaks, and then localizing the leaks to within 300m (Cai et al., 2023). Their approach was applied to a simulated water network and was able to detect 12 of 19 simulated leaks and to localize 8 out of 19.

Ravichandran et al. explored the use of acoustic sensor data to detect the presence of a leak in close proximity to a monitoring sensor (Ravichandran et al., 2021). The authors used power spectral

density based features and found that a strategy of training multiple gradient boosted trees and then creating an ensemble learner from these to be extremely effective, reaching an accuracy of 99.84%.

2.2.2 Case Studies and Survey Reports

2.2.2.1 Background on Case Studies and Survey Reports

Within the domain of pipeline management, making pipe replacement decisions is often considered an engineering discipline. As such, the professional judgment of an experienced engineer figures prominently. While this process may lack academic rigor, it is nevertheless common in practical application and thus merits description.

Beyond the foundational requirements of becoming a professional engineer (education and experience), there is a requirement for background information to be used by the engineer to inform their professional judgement. The background information comes primarily from a mixture of personal experience, peer discussion, and industry education (papers, articles, seminars, etc.). In the case of pipeline risk assessment, industry education consists largely of publications of survey results and case studies. Survey results often include highly aggregated information (such as average break rates) across multiple utilities. Case studies often include informal univariate analysis (charts and aggregations, usually without any formal statistical analysis) of data from a single utility or detailed analysis of the proximate cause of a single pipe break. These types of studies supplement an engineer's personal experience, helping them to provide informed opinions. The ease of understanding and applying these studies likely contributes to their popularity in industry.

2.2.2.2 Prior Work on Case Studies and Survey Reports for Pipe Diagnostics

A wide range of surveys and statistical studies have been conducted related to estimating the risk of pipeline failure. These have generally focused on a single parameter at a time, either as a categorial variable (such as pipe material) or a linear correlation (such as age).

Different types of pipe have been found to exhibit different mechanisms and rates of failure. Table 4 illustrates the findings of a study by the National Research Council of Canada that found the following average break rates per 100 km per year across 1992 and 1993 on different materials (Rajani et al., 1993).

Table 4: Water main break rates for Canadian pipes of a variety of materials, based on data found in (Rajani et al., 1993)

Material	Length (km)	1992 Breaks		1993 Breaks		1 Year Change	
		Total	/100 km	Total	/100 km	Rate	As %
Cast Iron	8769.9	3078	35.1	3219	36.7	1.6	4.6%
Ductile Iron	4237.5	394	9.3	415	9.8	0.5	5.4%
Asbestos Cement	2105.4	114	5.4	128	6.1	0.7	13.0%
PVC	1818	16	0.9	9	0.5	-0.4	-44.4%
PCCP	623.2	3	0.5	5	0.8	0.3	60.0%
Total	**17554**	**3605**	**20.5**	**3776**	**21.5**	**1**	**4.9%**

Cast iron pipes exhibit the highest rates of failure. Many cast iron pipes have no corrosion protection applied to their inside and/or outside diameter, leaving them prone to rust. The material is also brittle and prone to cracking. There have also been periods, particularly in the 1930s, when a material called leadite was used for packing the joints of cast iron water mains (Makar et al., 2001). This material has a different coefficient of thermal expansion than the iron itself, causing added stress near the joints during temperature cycles, which is thought to result in cracking of the bell end of the pipe.

Ductile iron pipes share some failure mechanisms with cast iron pipes, with two key differences. First, the material is less brittle, making these pipes capable of supporting small, non-surfacing leaks for long periods without bursting. Second, newer ductile iron pipes have inner linings and external coatings applied to reduce the rates of corrosion.

Steel, plastic, and PVC pipes are all less prone to corrosion. These materials are all very flexible, leaving them less prone to brittle failure as well. Many failures in these pipes can be traced to the joints, which can be the weak points in the pipe. Water leaking from the joints can also impact the surrounding soil, creating voids (loss of support) or frost heaves that cause unplanned stress on the pipe.

> Calgary inspectors have learned the hard way that permitting deviations of even a few degrees on main over 200mm offer a slight but measurable risk of leakage at the gasket, which then begins a process of erosion-corrosion of the gasket, the bedding, and even the

> pipe itself. The leak only increases with time until finally a major failure can occur.
>
> (Brander, 2004, p. 8)

Recorded failures in PCCP happen at quite low rates for two reasons. First, the pipe barrel does not exhibit either of the gradual failure mechanisms (cracks, or corrosion pinholes) shown by ferrous pipes. These pipes tend to burst catastrophically rather than leaking. When leakage does occur, it is often at the joints. The exception to this is Lined Cylinder Pipe, a type of PCCP in which the prestressing wires are placed directly on the steel cylinder, where leakage from corrosion of the steel cylinder has been suggested as a failure mechanism (Erbay et al., 2007). Second, leaks on PCCP are difficult to detect using conventional acoustic leak detection approaches. More costly techniques are required, such as inline leak location or trunk main correlators, which are used less frequently (Laven & Lambert, 2012). Hence a leak on PCCP is more likely to go undetected than on other materials.

Diameter is another well-known predictive factor. Larger diameter pipes have been found to exhibit lower rates of failure than small diameter pipes (Sundahl, 1996). There are two known mechanisms for this. First, larger diameter pipes generally employ thicker pipe walls. This means that the same design safety factor (as a ratio) will result in a greater excess material thickness. Simply put, more of the pipe wall can corrode before reaching the critical thickness. The greater pipe wall strength also leaves it more resistant to stresses for which it was not designed, such as third-party damage (e.g., being struck during excavation) and torque applied during installation. Second, leaks on these mains are more likely to go undetected. Large diameter mains are typically buried more deeply, hence a greater portion of leaks do not surface. Sound attenuates more quickly on large diameter mains, and fittings for listening are placed less frequently, making acoustic leak detection methods less effective (Jones & Laven, 2008). These two factors both contribute to the lower rate of measured failures on large diameter mains.

The impact of age has also been widely investigated. While many failure mechanisms operate over long periods, the relationship between age and break rate is quite complex. Pipe materials have changed over time. The oldest material (Cast Iron) generally exhibits the highest failure rates; however, it is not immediately clear whether this is due to age or material. Similarly, PVC is the newest material widely used and exhibits the lowest failure rate; however, the age and material factors are highly correlated and difficult to decouple. Even within a given material type, there is

complexity. Material quality and installation practices have generally improved over time. For example, the first generation of Ductile Iron pipe lacked appropriate corrosion protection, resulting in high break rates. The "bathtub curve" of asset life also comes into play, as material or installation defects may surface early in the pipe's life. Survival bias also plays a role, as over poorly performing pipes (or cohorts) may be replaced early, leaving behind only those whose material, installation, and soil environment will allow them to operate for many decades without issues. The author has personally seen >100-year-old cast iron pipe that looked nearly new when coming out of the ground. The net result can be a rather complex relationship between age and break rate, as illustrated in Figure 3, as found in (Stone et al., 2002) and showing data from (Sundahl, 1996).

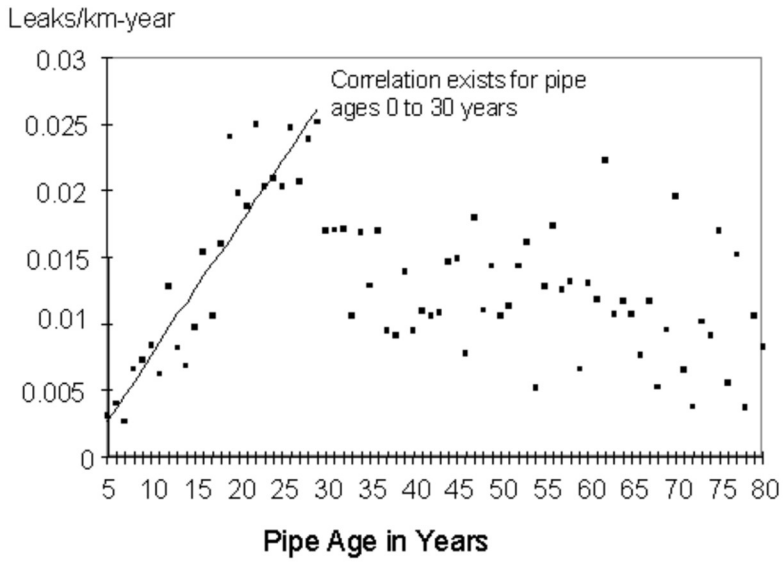


Figure 3: Failure rates by age for cast iron pipe in Malmo, Sweden, over five years (Sundahl, 1996).

Traffic loading has also been shown to be a predictor of failure rate, as shown in Table 4 from Stone et al. (2002), who cite Eisenbeis et al. (2000) as the original source of the data. It is not entirely clear whether it is the added weight, the variability of the weight, or the vibrations that cause this.

Table 5: Relative failures rates for water mains in high traffic and low traffic areas in France (Stone et al., 2002).

Relative Failure Rate = [h(high traffic)] / [h(low traffic)]	Bordeaux (GCI, 1st fail.)	Charente M. (GCI, 1st fail.)	Suburb of Paris (GCI, 1st fail.)
	2.30	3.00	1.77

Internal pressures also play a prominent role. Higher pressures have been found to be predictive of the rate of failure formation (Lambert, 1994). Variations in internal pressure, such as in systems with intermittent water pressure, can induce material fatigue due to cyclical stresses; systems experiencing pressure cycles have also shown higher rates of failures (Rezaei et al., 2015). Finally, pressure transients (also known as water hammers) have been shown in many cases to be the immediate trigger of a failure (Romer et al., 2007), and the frequency and severity of pressure transients have also been quantitatively shown to be predictive of failure rates (Rezaei et al., 2015).

2.2.3 Survival Analysis

Survival analysis is an engineering concept widely used to predict asset failures, with numerous studies available in the literature describing its application to water pipe diagnosis.

2.2.3.1 Background on Survival Analysis

Survival analysis is a branch of statistics concerned with estimating the expected time until a failure event. These methods employ a selection of functions to describe the expected lifespan of members of a population. Central to these is a survival function S, which is denoted as:

$$S(t) = Pr(T > t) \tag{2}$$

Where:

- t is a time of interest
- T is the time of the failure event
- $Pr()$ is the probability

A survival function requires clear definition of the start time (i.e., $t = 0$). In most cases all members of the population are considered to survive at time zero (i.e., $S(0) = 1$); however, this is not strictly required.

It is, however, required that the survival function be non-increasing. In other words, it must be the case that once a member of the population has failed, it remains failed in the future, although it may be replaced. This requirement is expressed as:

$$S(t_2) \leq S(t_1) \text{ if } t_2 \geq t_1$$

(3)

The complement of the survival function $S(t)$ is the lifetime distribution function $F(t)$. It can be thought of as a cumulative failure distribution function. It is expressed as:

$$F(t) = Pr(T \leq t)$$

$$F(t) = 1 - S(t)$$

(4)

In cases where $F(t)$ and $S(t)$ are differentiable, then the survival event density $s(t)$ and failure event density $f(t)$ can be defined as follows:

$$s(t) = \frac{d}{dt}S(t)$$

$$f(t) = \frac{d}{dt}F(t)$$

(5)

These properties represent the probability density function of a sample failing at (or surviving through) time t. They are related as:

$$s(t) = -f(t)$$

(6)

The final key concept is the hazard function $h(t)$, also referred to as the conditional hazard function. This function represents the chances that a particular member of the population which survived up to time t will fail at time t. This is formally defined as the failure rate at time t, conditional upon having survived up to time t:

$$\mathrm{h}(t) = \lim_{\delta \mathrm{t} \to 0} \frac{Pr(t \leq T < t + \delta \mathrm{t})}{\delta \mathrm{t} \cdot S(t)}$$

$$h(t) = \frac{f(t)}{S(t)}$$

(7)

In many practical cases, data is available only in discrete time intervals (days, weeks, months, years, etc.). For these situations, the hazard function $h(t)$ can be approximated by the probability that a failure event will occur to a given member of the population at any point during the time interval beginning at t, conditional upon it having survived to the beginning of time interval t. The remaining curves $S(t)$, $s(t)$, $F(t)$, and $f(t)$ can be approximated in analogous manners.

Note that the hazard function, survival function, and failure functions all have simple relationships. Once any one is known, the others can be derived either analytically or numerically. Hence, fitting any one to data is sufficient to generate all of these curves.

Once these curves are fit, two commonly used derived quantities are the mean residual lifetime and the mean time to failure. The mean residual lifetime is the expectation value of future lifetime, given survival to age t_0. Where $t_0=0$, this becomes the mean time to failure (sometimes called the life expectancy at birth). Both are analytically solvable for many of the distributions commonly used and can be approximated numerically for nonparametric formulations (often ignoring the impact of long-tailed survivorship, which may be neglected in nonparametric models due to Right Censoring in data). Also of interest is the median residual lifetime and the median time to failure. These represent the times at which half of the population is expected to have failed. These median quantities are also analytically solvable for many distributions and can often be solved numerically for nonparametric formulations (provided that at least half of the study population experiences the failure event during the study timeframe).

2.2.3.2 Survival Curves Applied to a Cohort

Most survival curves are applied to a population as a whole. While the estimator for the curves may be parametric (e.g., Weibull curves) or nonparametric (e.g., Kaplan Meier curves), many are fit on a single parameter (time) and operate on an entire population.

There are numerous benefits to fitting survival curves to an entire population or cohort. The resultant curves provide a succinct summary of the expected behavior of a population over time and facilitate planning and decision making on the basis of these populations.

Parametric curves have several distinct advantages, thanks to the analytics expression of the distributions. These curves can often be used in conjunction with cost information for (failures and replacement) to identify analytical solutions to the optimal lifespan of members of the population.

Parametric curves can also be used to extrapolate to ages beyond what is included in the data to which they are fit.

Nonparametric curves offer their own advantages. Most importantly, they do not rely on assumptions as to the relationship between the failure rate and time (e.g., assuming exponential growth of the hazard rate). They are also more robust against outliers and noise in the training data. They often, however, require more data to fit the curves and, generally, cannot extrapolate to ages not present in the training data.

2.2.3.3 Survival Analysis Applied to Individuals

As larger data sets have become available, survival analysis has extended to take advantage of more predictive variables (sometimes called covariates).

The Cox Proportional Hazard model is a common approach to estimating the impact of factors other than time (covariates) on the hazard rate. It works from the assumption that a given covariate is multiplicatively related to the hazard function. This is expressed as:

$$h(t \mid \boldsymbol{X}_i) = h_0(t) \cdot e^{B_1 \boldsymbol{X}_{i1} + B_2 \boldsymbol{X}_{i2} + \cdots + B_p \boldsymbol{X}_{ip}} \quad \textbf{(8)}$$

Where:

- $h_0(t)$ is the baseline hazard function at time t for the entire population
- $\boldsymbol{X}_i$ is a vector of p covariates for sample number i, with entries $\boldsymbol{X}_{i1}$ through $\boldsymbol{X}_{ip}$
- $B_1, B_2, \ldots B_p$ are the coefficients applied to the p covariates, which are fit to data

Once a Cox Proportional Hazard model has been fit to data, the coefficients provide a clear indicator of how the value of a given covariate impacts the risk of failure. Different types of proportional hazard models are often referred to, based on the assumed shape of the baseline hazard function. For example, if a Weibull distribution is used for $h_0(t)$, it may be referred to as a Weibull Proportional Hazard model.

The Cox Proportional Hazard approach can be considered a parametric approach to inclusion of covariates, as it assumes a multiplicative relationship to the baseline hazard.

Nonparametric approaches to incorporating additional covariates also exist. Many of these rely on subdividing the population via tree structures. This can be done by successively dividing the population across one covariate at a time, creating a tree structure with separate populations at each leaf. A separate survival curve can then be fit to each leaf in the tree, creating a Survival Tree. Various strategies for dividing the population are available. The resulting conditional survival function $S(t \mid \boldsymbol{x})$ and conditional hazard function $h(t \mid \boldsymbol{x})$ can be smoothed out by fitting multiple Survival Trees, each on a randomly selected subset of the data. This approach, known as Survival Random Forests, has become common in recent years.

2.2.3.4 The Challenge of Censored and Truncated Data

In the ideal scenario for survival curve fitting, the entire lifespan of each member of the population is present in the records. In practice, some data is often unavailable, either due to events happening outside the study period (Censored) or samples being missing altogether due to the timing of events (Truncated). This data unavailability can be grouped into four categories (Turkson et al., 2021).

Right Censoring: It is known that a member of the population survived at least until some time l (the lower limit on the lift of that member of the population), but not precisely when. This often occurs when the failure occurred after the end of the data recording period.

Left Censoring: It is known that a member of the population failed before some time u (the upper limit on the life of that member of the population), but not precisely when. This often occurs when the failure occurred before the beginning of the data recording period.

Interval Censoring: It is known that a member of the population failed between time l (the lower limit on the lifespan) and u (the upper limit on the lifespan), but not precisely when. This often occurs when checks for the failure condition happen periodically, such as testing for presence of a disease.

Truncation: When members of the population with a lifespan less than some threshold are simply not observed. This often occurs when some members of the population experienced failure events prior to the start of the data recording period and are consequently not present in the study.

2.2.3.5 Prior Work on Fitting Survival Curves to Pipe Data

Failure curves have a long history of application in pipe diagnosis. A brief review of examples from the literature is provided in this section. First, however, some challenges must be addressed.

2.2.3.5.1 Challenges With Fitting Survival Curves to Pipe Data

There are four broad challenges in fitting survival curves to pipe data. The first is the ambiguity of what constitutes "failure" and "survival" of a pipe. The second is the problem of data censoring and truncation. The third is the challenge of selecting appropriate assumptions for fitting parametric models. The fourth challenge is finding a formulation that is directly applicable in practice.

The foundational assumption of survival curves is that a given member of the population can only fail once. It survives until the point of failure, at which point it is no longer a survivor. This is not strictly true of pipes at any level of granularity (stick, segment, or cohort). A stick of pipe can experience a break, be patched and put back into service, and then break again at a later date. Likewise, a pipe segment can break multiple times at different locations along its length, or even multiple times at the same location (if that location has been repaired). This ambiguity leaves several possible interpretations of a "failure event" in survival analysis:

- **The pipe is replaced**. This is clear and unambiguous and fits well with the definition of a failure event in Survival Analysis. It is, however, less useful, as the survival curves end up describing human decisions and utility policies rather than physical phenomena.
- **The pipe experiences its first break**. While clear and unambiguous, this is both of limited practical application and difficult to apply. It has the benefit of allowing calculation of a Mean Time to Failure for new pipes, which is an intuitive metric. However, most utilities accept a certain rate of water main failures as inevitable. Replacing pipes before their first failure is not generally their goal; in fact, many have a policy not to replace a pipe before it has failed at least once. The metric also provides limited insight into the behavior of pipes that have failed at least once, which are often the main area of focus for replacement programs. Fitting curves based on this definition is also difficult. The issues of Left Censoring and Right Censoring are both highly prevalent. Left Censoring is prevalent because most utilities do not have failure records stretching back nearly as far as the ages of their oldest pipes. Right censoring is prevalent because most pipes in their databases have not yet experienced a failure.
- **The pipe experiences its second break**. A relatively common policy is to replace a pipe after it experiences its second break. This definition shares some of the benefits and drawbacks of each of the two definitions above.

- **The pipe experiences its <u>next</u> break**. This resolves the problem of Left Censoring, as it allows for an arbitrary choice of the t=0 time. The t=0 time is often set at the time of the most recent break. This definition is still, however, prone to the Right Censoring and Truncation problems.
- **The pipe's expected rate of failures per km exceeds a threshold** (e.g., 20 breaks / 100 km / year). This is another replacement policy used by some utilities, as the average rate of breaks in the network (usually expressed per 100 km per year, or per 100 miles per year) is a common performance metric used to measure both customer service and the state of good repair of the network. It is not, however, a practical definition for fitting a survival curve since it requires a preexisting model for estimating the future likelihood of failure. This is equivalent to the hazard function itself, yielding a circular definition.

Each of the definitions above, except for exceeding an expected rate of failures, suffers from the second group of problems: Censoring and Truncation.

Right Censoring is a common challenge in fitting survival curves to situations where study durations are less than the maximum lifetime of the subjects. This is the case for pipe failure studies, where utility records typically span only a few decades, but pipes can remain in service for a century or more. There are four common methods for handling Right Censored data. The first is simply to discard the right-censored records. The second is imputation of the missing data, where a rule or model is used to fill in the missing records. Third is dichotomizing the data, where a classifier is first fit to separate those that did and did not experience a failure, and then fitting a survival curve to only the non-censored data. Fourth and most widely used is likelihood-based methods, which adjust mathematically for whether a sample was censored or not; this method is used by Kaplain-Meier estimators and by Cox Regression (Turkson et al., 2021).

Left Censoring in pipe failure records warrants deeper discussion. In a conventional Left Censoring scenario, it is known which members of the population failed prior to the study start time, but not when. An example of conventional Left Censoring would be leak detection on pipes, wherein a leak detection survey may discover a non-surfacing leak on the pipe which formed at some unknown time in the pipe's history. In pipe break records, however, another type of Left Censoring occurs, where it is simply not known whether a pipe has previously failed or not. This most commonly occurs when unknown breaks have occurred prior to commencement of systematic break

recording. This problem would not be present in an ideal scenario where a utility has recorded all pipe failures and pipe replacements since the installation of its first pipe. The reality is that recording pipe breaks is a relatively recent trend in industry. Many utilities do not practice it. Those that do generally have limited records. In the case of the utilities participating in this study, the commencement date of their records ranges from 1960 to 2010. This likely means that many breaks that occurred early in the lifetimes of some pipes are unrecorded. There is no definitive way to know which pipes have and have not previously failed. Consequently, most methods of accounting for Left Censoring, such as removing Left Censored records from the training set, cannot generally be applied. Such unknown missing records can lead to a bias in the survival curve tending to overestimate survival rates. The dichotomization method has, however, been applied successfully. To account for previously unrecorded leaks, a model to predict which pipes had breaks prior to commencement of recordkeeping can be created followed by a second model to estimate when. This approach has been shown to reduce the bias introduced by left censoring (Xu & Sinha, 2021).

Interval Censoring is not a widely documented problem for pipe failure records.

Truncation is also a common challenge in pipe failure records. Many utilities keep records of their current installed pipe inventory only. Once a pipe is removed from service, it is expunged from their records. This is a form of Truncation, whereby pipes which were removed from service prior to the study date are simply not present in the database. It acts as a form of selection bias, whereby pipes which failed early may simply not appear in the data. The absence of records for pipes which experienced premature failure may cause a bias in the survival curve tending to overestimate survival rates. Fortunately, many data sets (including the majority used in this study) also include records of pipes that had been removed from service, in which case no Truncation is expected.

The third group of problems relates to the challenging of selecting appropriate assumptions in fitting parametric and semi-parametric (where the relationship with time is parametric, but with other covariates is nonparametric) models. By their nature, parametric models assume some form for the relationship between time, the covariates, and the failure rate. Particularly common is assuming a specific form of the relationship between time and the failure rate, such as (for example) an exponential increase in the hazard rate with time or an exponential decrease in the mean time between failures with age. While not strictly required, most implementations of parametric or semi-parametric models choose a baseline hazard function which assumes a monotonic relationship between the

hazard rate and time for a given cohort of pipes. This relationship is shown in Chapters 4.2.4 and 5.4.2 to be highly questionable. Likewise, proportional hazard models assume that each covariate impacts failure risk consistently regardless of the value of other covariates. This too is a questionable assumption with respect to pipes, as shown in Chapter 4.2.2 with the interaction between pipe material and diameter and in Chapter 4.2.4 with pipe material and age.

The fourth and final challenge relates to finding a formulation of the survival curve that is directly applicable in practice. Fitting a survival curve requires choosing a study period, which in practical terms generally means the duration of pipe break records provided by utilities contributing data. Many of the resulting metrics, such as the mean time to next failure, would be applicable to other similar study periods. However, given the changing nature of pipe manufacturing and installation over time, it is not clear whether pipes will behave similarly during a future period (for example, starting 30 years later than the original study period). Holding back the most recent study data from the curve fitting process and using this to test the curve would provide this confirmation; however, this is not common practice in survival curve fitting. Similarly challenging is that many problem formulations require a long history of record-keeping (specifically when the most recent break occurred on each pipe) to be effectively applied.

2.2.3.5.2 History of Fitting Survival Curves to Pipe Data

Here are brief summaries of several studies related to survival analysis or survival curves in the context of water pipe break prediction:

Le Gat and Eisenbeis applied a Weibull Proportional Hazard Model on both long duration (20 to 50 years) and short duration (5 years) maintenance records (Le Gat & Eisenbeis, 2000). The study found that with appropriate both left and right truncated data and appropriate grouping of cohorts, the short duration data was sufficient to fit the survival curves effectively.

Scheidegger et al. explored the impact of truncation on pipe failure models (Scheidegger et al., 2013). They proposed an extension to these methods to consider absent data from replaced pipes. Their further work included a review of statistical failure models applied to water distribution pipes prior to 2015 (Scheidegger et al., 2015).

Xu and Sinha discussed the overlooked issue of left truncation in many studies. They highlighted that this leads to systematic bias in survival analysis models, affecting the scale and shape of survival

curves and changing Mean Time To Failure (MTTF) estimates (Xu & Sinha, 2020). In particular, they noted that this bias leads to underestimation of failure rates and overestimation of the time to next break. They later proposed integrating an Artificial Neural Network (ANN) imputation method with Weibull proportional hazard survival analysis to calibrate the survival curve and reduce bias in MTTF estimation due to left truncation, reducing bias from 14.3% to 2.1% (Xu & Sinha, 2021).

Phan et al. proposed an approach to managing water main breaks using risk-based decision making, combining machine learning and survival analysis (Phan et al., 2019).

Rahbaralam et al. employed machine learning algorithms and a Cox proportional hazard survival analysis model to predict water main failures in Barcelona, evaluating the models with various metrics including AUC and Matthews' Correlation Coefficient (Rahbaralam et al., 2007).

The use of survival trees, survival tree ensembles, and survival random forests for this application was introduced in a master's thesis (Oliveira, 2019). This thesis further introduced the concept of ranking the pipe segments and plotting the percentage of the network selected for replacement versus the percentage of failures avoided if it were replaced, describing this as a cumulative gain curve, the use of a concordance index for measuring overall performance, and the use of a metric (such as the Brier score) to measure the absolute accuracy of failure probability estimates.

Snider and McBean published several recent studies on this topic. They addressed the issue of censored data in water utility datasets, where many pipes in service have never experienced a break. They noted that traditional survival analysis models like Cox proportional hazard models can handle censored data, unlike many machine learning models (Snider & McBean, 2020b). They also compared a machine learning model (XGBoost) with Weibull proportional hazard survival analysis for predicting time to next failure, finding that the XGBoost model underpredicted time to next break due to its inability to include censored events (Snider & McBean, 2020a). They further discussed the use of survival machine learning models like Random Survival Forest, which incorporate censored data and model complex relationships between variables (Snider & McBean, 2021).

Recent work on fitting survival curves to data continues to explore the use of random survival forests (Daulat et al., 2024). Random survival forests can be considered both survival curve fitting and a machine learning approach. Prior work in this area will also be described in Chapter 2.3.2.

2.2.4 Failure Risk Assessment

The concept of failure risk prediction is prominent in engineering approaches to asset management. This section provides a synopsis of studies from the literature which have applied the technique to water pipe diagnosis.

2.2.4.1 Background on Failure Risk Assessment

Failure risk prediction deals with the same fundamental problem as survival analysis but approaches it from the opposite direction. Rather than focusing on the question of how long it will be until a member of the population fails, it focuses on estimating the risk of failure in a given time period.

Failure risk assessment generally breaks down the problem into two dimensions: the likelihood (or probability) of a failure occurring, and the consequences (or severity) of a failure if it does occur. Both dimensions can be handled qualitatively (such as with risk categories) or quantitatively (such as in failure probability estimation). Similarly, the joint consideration of likelihood and consequences of failure can also be handled qualitatively (such as with a risk matrix) or quantitatively. When joint risk consideration is handled quantitatively, a common approach is to calculate the expectation value of the consequences from a given risk:

$$RoF(t) = CoF(t) \cdot PoF(t) \tag{9}$$

Where:

- t = A time period of interest, such as a particular year
- $RoF(t)$ = Risk of Failure during time period t
- $CoF(t)$ = Consequence of Failure during time period t
- $PoF(t)$ = Probability of Failure during time period t

In many applications, a range of different types of failures are possible, each with different probability and consequences of failure. In such cases, the total risk is often expressed as a summation over these different types of risk.

$$RoF(t) = \sum_i RoF_i(t)$$

$$\mathrm{RoF}(t) = \sum_i CoF_i(t) \cdot PoF_i(t)$$

(10)

Where:

- RoF_i, CoF_i, and PoF_i are the Risk, Consequence, and Probability of the i^{th} failure of type.

The expression of the aggregate risk over a population comes in a similar form. Rather than summing over types of risk, the same summation is applied over the different members of a population. Where the population is homogeneous, this works out to a simple multiplication by the total size of the population.

$$\mathrm{RoF}(t) = \sum_{n=1}^{N} Cof_n(t) \cdot PoF_n(t)$$

$$RoF(t) = N \cdot CoF(t) \cdot PoF(t)$$

(11)

Where:

- RoF_n, CoF_n, and PoF_n are the Risk, Consequence, and Probability of a failure of the n^{th} member of the population
- N is the total size of the population

It should be noted that, while the terms "Probability of Failure" and "Likelihood of Failure" are commonly used, strictly speaking this should be the expectation value of the number of failures. The two are identical in the case where a particular failure can only occur once; however, in some cases, it is possible for the same member of the population to fail multiple times within the same time unit t.

This study is primarily focused on the PoF term. A range of methods exist for the estimation of this term. Several of these methods and their history of use for pipe failure risk assessment are described in the sections below.

2.2.4.2 Failure Prediction Applied to a Pipe Cohort

Early approaches to pipe failure probability estimation generally considered large populations of pipes together. These approaches generally involve first defining cohorts (i.e., separating the population into groups across one or more dimensions) and then predicting the number of failures within a given cohort and a given period.

Separating the population into groups (called segmentation) requires selection of both the dimensions (variables) upon which to divide the population and the thresholds (for numerical variables) or aggregations (for categorical variables) that will be used to separate the groups. This process can be done analytically (via segmentation techniques, such as K-Means) or manually. When performed manually, exploratory data analysis and statistical techniques are generally applied to select the variables for use in defining the cohorts and then to confirm that the resultant groups are appropriate.

Predicting the number of failures within a group is generally accomplished via regression analysis. Various types of regression analysis are available, varying in their complexity. Complexity can be increased across two dimensions: the number of predictor variables used and the assumptions made as to the form of the relationships assumed among the variables.

2.2.4.2.1 Segmentation Approaches

Analytical approaches to segmentation are generally formulated as optimization problems. These seek to minimize some cost function. Perhaps the best known segmentation algorithm is K-Means. This algorithm aims to identify central points for K groups in a manner that minimized the average distance between each member of the population and its assigned group central point. While various distance measures are possible, Euclidean Distance is the most common.

One practical challenge with K-Means segmentation is that the results can be challenging to understand and explain. As the separations between classes exist across multiple dimensions of the data, it is often the case that no clear definition of the barrier (e.g., “diameter > 150 mm” in the case of pipe populations) is possible. This is particularly true when the data is subject to dimensionality reduction (such as principal component analysis) prior to application of the K-Means algorithm. Consequently, the segments are generally “profiled” across additional dimensions of the data and given qualitative labels which are easy to understand. A clear example of this in the domain of water

pipelines is the distinction between classes of pipes: “customer connections,” “distribution mains,” “trunk mains,” and “transmission mains.” While there are no strict demarcation lines between the segments, the concepts are understandable. For example, “distribution mains” generally describes water pipes that are of relatively small diameter and short length, constructed with low-cost materials that are easy to tap, but not directly supplying a single customer with water.

In practice, K-Means is often applied hierarchically as a means of creating more explainable segments, first segmenting into K groups, and then further segmenting some or all of these K groups into subsegments. Selecting a very small number of dimensions for each level of segmentation can create segments which are easily explained by their decision parameters (e.g., “Plastic pipes less than 75m in length” for pipelines).

Manual approaches for constructing segments often yield the most explainable segments, making them quite common in practice. In such cases it is common practice to begin with exploratory data analysis, considering predictor variables individually (or in pairs) and checking for relationships with the target variables.

It is the author’s professional observation that manual rules for segmenting the pipe population into cohorts are quite common in the pipeline industry. It is not clear whether this is due to a desire for explainable segments, the small number of predictive variables available, or the existing research regarding the relationships between these and the target variable.

2.2.4.2.2 Regression Approaches

Once cohorts have been defined, predicting the number of future breaks within a cohort can be treated as a regression problem. Regression models involve estimating some quantity y based on a vector of predictive variables $\boldsymbol{x}$ using an estimator function. While a time variable t can be contained within $\boldsymbol{x}$, in the case of regression problems that attempt to estimate a future value, t is often shown separately.

$$\hat{y} = f(\boldsymbol{x})$$

$$\hat{y} = f(t, \boldsymbol{x})$$

(12)

Where:

- $\hat{y}$ is the estimated value of target (dependent) variable y

- $\boldsymbol{x}$ is a vector of predictive variables
- t is a time period of interest
- $f()$ is an estimator; a function which estimates the value of ŷ at time t using $\boldsymbol{x}$

In the case of pipe failure risk assessment, the target variable y is generally the number of breaks within a given cohort in a given time period t. The predictive variables used for $\boldsymbol{x}$ generally consist of the variables used to define the cohort, concatenated with several statistical measures on the cohort.

$$B_m(t) = f(t, [cohort\_identity; cohort\_statistics]) \quad \textbf{(13)}$$

Where:

- t is the time period of interest
- $B_m(t)$ is the number of breaks in the m^{th} cohort during time period t
- *cohort_identity* is a vector of the variable values used to define the cohort, such as pipe diameter, pipe material, or geographic neighborhood
- *cohort_statistics* is a collection of aggregate statistics of the pipes in the cohort, such as average age, total length, or the failure rates from prior time periods

A wide range of estimators can be used for the function $f(t, \boldsymbol{x})$. They can be broadly classified into point-in-time models, which use information about the current state of the system in $\boldsymbol{x}$, and time series regression models (also called autoregressive models), which use past values of the target variable in $\boldsymbol{x}$. Some models are able to consider both current state and historical information jointly in $\boldsymbol{x}$.

Perhaps the simplest point-in-time model is linear regression, which uses only a single variable for x, and assumes a linear relationship between x and y. Complexity can be introduced in either of two ways: increasing the number of variables in $\boldsymbol{x}$ or using estimators $f(\boldsymbol{x})$ which make fewer (or weaker) assumptions about the relationship between variables.

The simplest time series model is the trivial identity model, whereby each sample in the series is predicted to be the same as the previous time interval: $f(t) = y(t\text{-}1)$. As with the point in time models, complexity can be added both by increasing the number of variables considered (e.g., $f(t)$ =

MEAN($y(t\text{-}1)$, $y(t\text{-}2)$, $y(t\text{-}3)$) or by making fewer (or weaker) assumptions about the relationship between $f(t)$ and the values of y during previous time intervals.

This distinction is not strict, as both point-in-time models and time series models can be adapted to blur the distinction between the two. Point-in-time models can include lag variables in $\boldsymbol{x}$, which provide information about the historical state of the system. Certain autoregressive models, such as Markov Chains and Recurrent Neural Networks, consider both past values and current state variables jointly when making a prediction of the next value.

2.2.4.3 Failure Prediction Applied to Individual Pipes

When applied to an individual pipe, failure prediction can be formulated as a classification problem. The common approach is to classify pipes as breaking vs not breaking within a particular timeframe. The general formulation of this approach is to predict a class label y based on a vector of predictive variables $\boldsymbol{x}$, using a classification function.

$$\hat{y} = g(\boldsymbol{x}) \tag{14}$$

Where:

- $\hat{y}$ is the predicted class label
- $g(\boldsymbol{x})$ is a classification function, which returns the predicted class label

The classification function $g(\boldsymbol{x})$ can be formulated as seeking the label assignment that maximizes the value of a scoring function on that data:

$$g(\mathbf{x}) = \underset{i}{\operatorname{argmax}}\ \mathrm{f}(\boldsymbol{x}, y_i) \tag{15}$$

Where:

- $\boldsymbol{x}$ is a vector of numerical features
- y_i is the i^{th} possible label which could be assigned to the sample represented by $\boldsymbol{x}$
- $f(\boldsymbol{x}, y_i)$ is the scoring function, where label y_i is applied to feature vector $\boldsymbol{x}$

Many different scoring functions can be employed; however, a commonly used class of scoring function in machine learning models is to use a conditional probability estimate:

$$f(\boldsymbol{x}, y) = P(\, y \mid \boldsymbol{x}\,)$$

(16)

Where:

- $\boldsymbol{x}$ is a vector of numerical features
- y is a possible label which could be assigned to the sample represented by $\boldsymbol{x}$
- $P(\mathrm{y} \mid \boldsymbol{x})$ is the estimated probability that y is the correct label for the sample, given that the sample is represented by feature vector $\boldsymbol{x}$

This represents an estimate of the probability of the label y being correct, given the observed data vector $\boldsymbol{x}$. In the case of failure prediction, this is generally formulated as a binary classification problem, where only two classes (Failure = true or false) are possible, and with the time period t shown separately from the rest of $\boldsymbol{x}$. This simplifies the formulation to estimation of a single probability (that Failure = true):

$$f(t, \boldsymbol{x}) = P(\,\text{Failure occurs at time } t \mid \boldsymbol{x}\,)$$

$$f(t, \boldsymbol{x}) = PoF(\, t \mid \boldsymbol{x}\,)$$

(17)

Where:

- $\boldsymbol{x}$ is a vector of numerical features
- t is a time of interest
- $PoF(t \mid \boldsymbol{x})$ is the Probability of a Failure during time t for the sample described by $\boldsymbol{x}$

Note that in failure risk assessment, the Probability of Failure during time t is assumed to be conditional upon the sample having survived up to time t. As such, this formulation of the scoring function matches the Hazard Function $h(t \mid \boldsymbol{x})$ used in Survival Analysis, described in Chapter 2.2.3.3. Care should be taken not to confuse it with the failure event density, which is the joint probability of the sample surviving up to time t and also failing during time t.

2.2.4.3.1 Defining a Member of the Population and the Target Variable

Many descriptions of the problem, such as "pipe failure prediction," seem intuitively clear, but contain hidden ambiguity. This is a consequence of certain definitions being subjective and a result of utility record-keeping decisions. In particular, the following three questions must be answered:

- What constitutes a single pipe?
- What constitutes a break?
- What sampling frequency should be used across a pipe's lifetime?
- What should be each prediction's time window?

Answers to these questions are required to unambiguously define a member of the population for analysis and the target variable. Along with being unambiguous, for the model to be practically applicable, the definitions of a member of the population and the target variable should also match the manner in which decisions are made.

Ambiguity in what constitutes a member of the population (i.e., a pipe) for a given study can be particularly problematic. As noted in Chapter 2.1.1.1.2, the term *pipe* can refer to a stick, a segment, or a pipeline. If a utility were to take a model for predicting failure risk on a stick and use it to predict the probability of failure on an entire pipeline, this could result in failure risk estimates that are off by several orders of magnitude. Even where it is clear that a pipe refers to a segment, the precise rules for what constitutes a segment vary from utility to utility. This limits the transferability of models or failure rates from one utility to another.

To be practically applicable, the definition of a pipe should match the items upon which decisions are made. For example, if a utility aims to extend the life of a pipe by reinforcing individual sticks of pipe, then providing failure risk predictions on the basis of full pipe segments (which may include dozens or even hundreds of sticks) is not sufficiently granular to be practically applicable. Conversely, if a utility is making decisions regarding entire neighborhoods to undergo pipe replacement, failure risk predictions on a per-segment basis may be too granular.

Similar challenges exist in the definition of a break. As described in 2.1.1.3, pipe failures exist on a continuum from inconsequentially small leaks (e.g., joint seepage) through to catastrophic bursts. Each utility must set a threshold of severity which qualifies as a break and should be entered into its

records. These thresholds may be different from one utility to another. Many utilities have two distinct programs for repairing broken or leaking pipes: reactive and proactive. A reactive program requires reports of a water main break to be provided by the public, at which point a repair crew is usually dispatched immediately. These are generally considered breaks and are recorded in a maintenance system. Some, but not all, utilities also have a proactive program, which involves using inspections or monitoring systems to detect non-surfacing leaks. Where such programs exist, these records may or may not be entered into the break history system. Another point of variation is leaks on a meter connected to a water main; some utilities may consider this as a break on that main, whereas others may not.

The selection of sampling frequency and time window are also significant. While it is common for these two values to be the same (e.g., predicting once per year whether there will be a break in the upcoming one-year period), they can also be different (e.g., predicting once per month whether there will be a break in the upcoming twelve-month period). Where data granularity supports using a higher sampling frequency (with overlap between time periods), this approach can generate larger training and test sets from the same underlying data.

The choice of prediction time window is particularly significant for practical application purposes. This choice has two components: the duration (e.g., 6 months, 1 year, 5 years) and the lag if any (e.g., predict breaks in the one-year period starting 1 year from now). Often, predictions are made using pipe records for a single year in the future. In practical terms, the data from a prior year is often not available until several months into the next year, and predictions would be available even later than that. Decisions about which pipes to replace that year, however, would generally be made prior to the beginning of the year. This means that the predictions would only become available once they were no longer necessary. Introducing a lag or making predictions over a longer period can make the predictions easier to apply.

The choice of time window duration can also impact model performance. Failure prediction datasets are often unbalanced, with more negative (did not fail) records than positive (did fail) in any given year. Some machine learning algorithms require consideration for imbalanced data, such as stratified sampling or weighting samples to rebalance the classes. Lengthening the period for the target variable can reduce the degree of imbalance in the data set. Care must be taken, however, to

avoid or account for the same Right-Censorship problem present in Survival Analysis, as described in Chapter 2.2.3.4.

2.2.4.3.2 Selecting the Performance Metrics

A wide range of performance metrics are available for comparing the performance of different models. Some, such as Accuracy, Precision, and Recall, are dependent upon the threshold used to separate positive from negative classes. Others, such as Maximum F1 Score (often called the F1 Score), the Area Under the Curve for the Receiver Operating Characteristic (often called AUC), and the Cumulative Lift at a given percentage (often called Lift) are independent of this decision.

Selection of an appropriate performance metric is important for the practical use of the resulting model. A good metric will demonstrate to any skilled professional in the problem domain whether the particular model is appropriate for their application. Poor choices of performance metric can lead to confusion or even be deceptive. An example of a poor metric would be using Accuracy in the case of predicting whether a pipe will break in the ensuing year.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Number of Samples}} \quad \textbf{(18)}$$

The reason this is a poor choice is that the data will generally be highly imbalanced. Many studies have shown that a given pipe will have less than a 2% chance of failing in a given year. That means that the trivial classification rule "pipes do not break" will be more than 98% accurate. This may sound like excellent performance to a layperson, despite being a terrible model.

In the case of pipe failure prediction, a wide array of performance metrics have been reported in the literature. These include Accuracy, Precision, Recall, F1 Score, ROC Area Under the Curve, Concordance Index, and others (Barton et al., 2022a).

2.2.4.4 Prior Work on Failure Risk Assessment to Pipe Diagnostics

This section provides a brief review of studies in the literature of Failure Risk Assessment applied to pipelines.

In their comprehensive exploration, Clark et al. (2002) developed cost models for various aspects of water supply distribution systems. The paper focuses on estimating costs for construction,

expansion, rehabilitation, and repair of distribution system components. It presents equations for the installation of new pipes, trenching, embedment, and several unit operations necessary for new construction or replacement of distribution system components. Additionally, the study develops cost estimations for relining and rehabilitation methods. These models are intended to assist in making preliminary cost estimates for a wide range of repair and rehabilitation scenarios, providing a valuable tool for urban infrastructure management.

Vairavamoorthy and Ali developed a Dynamic Risk-Based Maintenance model for water distribution networks that evaluates both the probability of pipe failure and its consequences (Vairavamoorthy & Ali, 2005). This model employed a genetic algorithm and proved effective in determining optimal maintenance and rehabilitation strategies by balancing the costs of maintenance against the risks and consequences of pipe failures.

Giustolisi et al. proposed a methodology for assessing failure risk in water distribution systems, emphasizing the importance of understanding both the likelihood of failure and its consequences (Giustolisi et al., 2006). The methodology successfully quantified risks and helped prioritize pipes for maintenance and replacement.

De Marinis et al. introduced a risk-based decision support system for rehabilitating water distribution systems. The system combined hydraulic modeling with failure risk analysis and was effective in optimizing rehabilitation strategies, leading to cost savings and improved system reliability (De Marinis et al., 2009).

Fares and Zayed (2010) developed a hierarchical fuzzy expert system to assess the risk of water main failures, considering 16 risk factors. They categorized these factors into two groups: those leading to failure (deterioration factors) and those resulting from failure (consequence factors). The study found that the most significant influences on failure risk were pipe age, pipe material, and pipe breakage rate, offering a comprehensive risk assessment approach for water mains.

Kabir et al. proposed a Bayesian Belief Network model to prioritize metallic water mains and evaluate their failure risk. The model incorporated structural integrity, hydraulic capacity, water quality, and consequence factors (Kabir et al., 2015). They highlighted the flexibility of their model to include additional factors, providing a robust and adaptable framework for assessing water main failure risks.

Malm et al. conducted a risk-based Cost-Benefit Analysis for leakage reduction, comparing the costs and benefits of different alternatives over time, including uncertainty analysis. They found that reactive repairing, despite a high leakage rate, was more cost-effective for their study network compared to proactive pipe replacement, providing valuable insights into the economic aspects of water main break management (Malm et al., 2015).

Al-Zahrani et al. applied a fuzzy synthetic evaluation technique for risk-based prioritization of water main failures. This approach incorporated various risk factors and uncertainties to effectively prioritize maintenance and intervention efforts based on the assessed risk levels (Al-Zahrani et al., 2016).

Marzouk and Osama presented a coordinated maintenance method for road, water distribution, and wastewater distribution networks. They developed a deterioration model using a hierarchical fuzzy expert system, a risk model using fuzzy Monte Carlo simulation for probability of failure, and Analytic Hierarchy Process for consequences of failure (Marzouk & Osama, 2017). Their multi-objective optimization using a genetic algorithm focused on minimizing overall risk, maximizing service levels, optimizing asset conditions, and minimizing life cycle costs.

Phan et al. employed a risk assessment framework in a case study of water main in a water distribution network, using Weibull distribution for failure probability calculation. They used a fuzzy inference system to aggregate different types of failure consequences, which included impacts on the network's redundancy/vulnerability, water loss, rehabilitation costs, and public health (Phan et al., 2019). This approach allowed for the integration of diverse impact factors into a unified outcome, aiding in comprehensive risk assessment and prioritization.

Vishwakarma and Sinha used a fuzzy inference method to develop a quantitative risk matrix for assessing the consequences of water pipeline failures. This approach aimed to reduce subjectivity compared to semi-quantitative and qualitative matrices. The model encompassed various types of consequences such as economic, environmental, social impacts, and operational intelligence (Vishwakarma & Sinha, 2020). This provided a more comprehensive and objective framework for visualizing and comparing failure risks.

Balekelayi and Tesfamariam utilized an ordered weighted averaging technique to identify the criticality index of wastewater pipes in the City of Calgary, combining it with a dynamic deterioration model to determine operational risk (Balekelayi & Tesfamariam, 2021). The technique was

particularly useful in identifying critical pipes when hydrodynamic data were not available, with the aim of aiding municipalities in prioritizing inspection and replacement plans.

In recent years, numerous studies have been published on the use of machine learning to predict the failure risk of individual pipes. While also relevant to the topic of failure risk assessment, a review of these applications in the literature is saved for Chapter 2.3.2.

2.2.5 Decision Optimization

Asset replacement decisions can also be formulated as optimization problem, either to minimize costs or maximize yield. Pipeline replacement and construction decisions using this approach are described in the literature, using predictions of pipeline failure as part of the approach.

A general formulation of the costs associated with water main management is provided in (Neelakantan et al., 2008):

$$C = IC + BC \tag{19}$$

Where:

- C = Total Costs
- IC = Installation Costs
- BC = Break Costs

The formulas for IC and BC are provided on a present value basis and can be adapted to cover a single pipe:

$$IC = L * UIC(m, d) \tag{20}$$

Where:

- L = length
- UIC = Unit Installation Costs, which is a function of the design parameters
- m = material

- d = diameter

This formulation assumes that pipeline installation costs correlate linearly with the length of the pipeline. This is a reasonable assumption. Most pipeline replacement is handled via contracts issued via a public tender process. These processes generally use costs per unit distance as the units of measure for the contract (BCC Research Inc, 2016), which forces the cost to the utility to fit the above formulation (Neelakantan et al., 2008).

$$BC = \sum_{t=1}^{T} (1 + DR)^{-t} * (1 + IR)^{t} * CoF(d, m) * PoF(m, d, L, age_t, etc.) \quad \textbf{(21)}$$

Where:

- t = a year of interest, typically starting with the upcoming year (t=1)
- T = total number of years for analysis
- DR = Discount rate for future cashflows
- IR = Inflation rate for break repair costs
- m, d, and L are the material, diameter, and length of the pipe, respectively
- CoF = Cost of Failure to repair a break, a function of diameter and material
- PoF = Probability of Failure in year i, a function of material, diameter, the age of the pipe in year t, and other available data of the pipe segment which helps to predict failures.

The cost function need not be limited to direct financial costs to the utility. Other factors, such as external costs to the community and the cost of greenhouse gas emissions, can be considered jointly as part of a combined cost function. These can also be incorporated directly into the financial cost via estimated damage claims from third parties or an assumed future carbon price.

If we make the simplifying assumption that the inflation rate for break costs will match the discount rate, these two terms will cancel out, leaving the break cost equation as:

$$BC = \sum_{t=1}^{T} CoF(d, m) * PoF(m, d, L, age_t, etc.)$$

(22)

Worth noting is the similarity of this equation to the formulation of the Failure Risk Assessment equations described in Chapter 2.2.4.3, with both the Cost of Failure and the Probability of Failure terms taking a similar form.

Starting from these equations, various constrained optimization problems can be formulated. Examples of such questions related to the decisions outlined in Chapter 2.1.4 are:

- Select pipes to replace which minimize the total number of expected breaks in the next five years, subject to the constraint of the total cost of replacement being less than the available budget.
- Select the replacement age which offers the minimum average annual total cost for 150mm PVC pipe, assuming *IR*=0.04 and *DR*=0.06.
- Identify the minimum average annual costs for 100mm pipe of each pipe material, selecting the optimal replacement age for each per the formulation above.
- Select the pipe diameter for ductile iron pipe which offers the minimum total cost over a 100-year time horizon.

With estimators for both *CoF* and *PoF*, all of the above formulations (among many others) can be solved relatively easily, providing effective guidance for decisions on a variety of timescales.

2.2.6 Specific Problem Statement for This Study

This study aims to provide a single model that generalizes across utilities, applications, and time periods. More specifically, the model should:

Extrapolate well to a new utility that did not contribute data to the model calibration:

- Regardless of language, jargon, and units of measure
- Regardless of their administrative practices (i.e., their definitions of a “pipe” and a “break,” and the availability of their failure history data)

- Regardless of their local context (e.g., climate, rainfall, pipe manufacturers, etc.)

Be applicable to a variety of applications:

- To estimating future number of pipe breaks in a cohort of arbitrary definition.
- To prioritizing a utility's pipes for replacement in the near future.
- To projecting the implications of design policies on future break rates.
- To supporting existing engineering practices, such as survival analysis and failure risk estimation.

Extrapolate forward in time, to any period of interest:

- Providing far enough ahead to be used in typical decision-making cycles.
- Validated as extrapolating forward in time.
- Also able to provide forecasts for alternative time periods.

A model able that meets these criteria should facilitate practical answers to many of the questions posed in Chapter 2.1.4.

The specific formulation chosen for this study involves estimating the Probability of Failure terms found in Chapters 2.2.3, 2.2.4, and 2.2.5 present in Survival Analysis, Failure Risk Assessment, and Decision Optimization respectively. It requires two equations:

$$ENoF(t \mid \boldsymbol{x}) = c \cdot PoF(t \mid \boldsymbol{x})$$

$$PoF(t \mid \boldsymbol{x}) = f(t, \boldsymbol{x}) \quad \textbf{(23)}$$

Where:

- t is the first year of a five-year period
- $\boldsymbol{x}$ is a feature vector describing a single pipe
- $ENoF(t \mid \boldsymbol{x})$ is Expected Number of Failures in the five-year period beginning at t, for the pipe described by feature vector **x**

- $PoF(t \mid \boldsymbol{x})$ is Probability that a pipe described by feature vector **x** will experience one or more failures during the five-year period beginning at t
- c is a calibration constant to convert between probability of failure and number of failures
- $f(t, \boldsymbol{x})$ is an estimator function

The following restrictions are placed on the estimator function $f()$ and $\boldsymbol{x}$:

- $f(t, \mathbf{x})$ must provide results for any starting time t after the pipe's installation date
    - This allows the model to be used in both near and long-term applications.
- $f(t, \mathbf{x})$ must be effective when extended forward in time to future time periods
    - This ensures that the model can predict future breaks, not just past ones
    - A practical implication of this restriction is that x may not include any features which incorporate information from the future
    - A practical implication of this restriction is that $\boldsymbol{x}$ may include features which incorporate information from the past (e.g., past failures), but $f(t, \mathbf{x})$ must be effective even if these are limited or absent
- $f(t, \mathbf{x})$ must be effective when extended to a new utility
    - This ensures that the model can be used by new utilities not part of the study
    - A practical implication of this restriction is that $\boldsymbol{x}$ must incorporate data as provided by the utility, without undergoing utility-specific data cleansing prior to use

A model conforming to these restrictions should achieve all of the generalization requirements outlined herein.

The application of the calibration constant c to convert between Probability of Failure and Expected Number of Failures facilitates use of the same model both to prioritize pipes within a utility, and to estimate the number of future failures in an arbitrary cohort by simple aggregation of the estimates for all pipes within that cohort. By selecting a cohort that matches the entire pipe population of a utility, this method would also support budgeting applications.

The restriction requiring effectiveness even when failure history is absent ensures that the model can be used for long-term forecasting as well, simply by using a future value of t to test the model. It also facilitates understanding the implication of design decision, by modifying a design parameter (e.g., changing from Cast Iron to PVC for new 100mm distribution mains) and forecasting future breaks under each scenario.

Choosing a problem formulation that also appears in each of the Survival Analysis, Failure Risk Assessment, and Decision Optimization applications means that the outputs of this model can be applied directly in these disciplines.

A default five-year time horizon for forecasting has been selected to match the typical duration of utility master planning activities, within which pipe replacement prioritization decisions are generally made. This also has the benefit (vs a one-year forecast) of providing better balance between the "failed" and "did not fail" classes.

2.3 Machine Learning for Pipe Failure Prediction

The approach taken in this study is to train a machine learning model to act as an estimator $f(t, \mathbf{x})$ of the conditional probability of a pipe failing $PoF(t \mid \boldsymbol{x})$, using data contributed by water utilities from around the world. This section provides background on the applicability of machine learning for pipe failure prediction, a review of past approaches described in the literature, and the specific methodology used in this study.

2.3.1 Background on Machine Learning for Failure Prediction Problems

A very brief introduction to machine learning methods is described herein. Further details about the application of these concepts to pipeline diagnostics are provided in later sections.

Machine learning methods are broadly classified into supervised and unsupervised learning algorithms.

Supervised machine learning models predict the value of a target (dependent) variable from a set of input (independent) variables. Two types of supervised learning algorithms are classification (for categorical target variables) and regression (for numerical target variables). Some diagnostic problems can be directly addressed using supervised learning methods, such as classifying items as "will fail" or "will not fail" over a given time. These algorithms are the focus of this project.

Unsupervised machine learning models operate without a target (dependent) variable. These models also fit parameters to minimize a cost function on training data. The most common class of unsupervised machine learning models are clustering algorithms, such as K-Means, which classify members of a population into a number of groups in which the elements are similar to each other. An example of using these algorithms for pipeline diagnostics would be to define cohorts of pipes which behave similarly to one another.

The supervised machine learning concepts planned for use in this paper are introduced below.

2.3.1.1 Machine Learning for Classification Problems

Each member of the population is described by a vector of numerical values (features). A scoring function describes how well the model correctly classifies a set of examples (the training set) with the value of the target variable provided (the ground truth). Fitting a model (also called training the model) involves searching for the set of parameters which minimize the cost function across the training set.

For classification, the target variable consists of one or more classes (called labels) to which a member of the population may belong. The classification function g(x) can be formulated as seeking the label assignment which maximizes the value of a scoring function on that data:

$$g(\mathbf{x}, \boldsymbol{\theta}) = \underset{y}{\operatorname{argmax}}\ f(\boldsymbol{x}, y, \boldsymbol{\theta}) \tag{24}$$

Where:

- $\boldsymbol{x}$ is a vector of numerical features
- y is a possible label assignment
- $\boldsymbol{\theta}$ is a vector of parameters
- $f(\boldsymbol{x}, y, \boldsymbol{\theta})$ is the scoring function, as applied when label y is applied to data vector $\boldsymbol{x}$, using parameter values in $\boldsymbol{\theta}$

By convention, $f(\boldsymbol{x}, y, \boldsymbol{\theta})$ may also be negatively oriented (i.e., lower scores are better), and described as a cost function or a loss function. In such cases, $g(\boldsymbol{x}, \boldsymbol{\theta})$ will either seek the maximum of $-f(\boldsymbol{x}, y, \boldsymbol{\theta})$, or the minimum of $f(\boldsymbol{x}, y, \boldsymbol{\theta})$.

Many different scoring functions can be employed. A commonly used type of scoring function in machine learning models is those which provide a conditional probability estimate:

$$f(\boldsymbol{x}, y, \boldsymbol{\theta}) = P(y \mid \boldsymbol{x}, \boldsymbol{\theta}) \tag{25}$$

Where:

- $\boldsymbol{x}$ is a vector of numerical features
- y is a possible label assignment
- $\boldsymbol{\theta}$ is a vector of parameters
- $P(y \mid \boldsymbol{x}, \boldsymbol{\theta})$ is probability that y is the correct label of the sample, given that it is described by feature vector $\boldsymbol{x}$ and the parameter values in $\boldsymbol{\theta}$ are used

This represents an estimate of the probability of the label y being correct, given the observed data vector $\boldsymbol{x}$ and the set of parameters $\boldsymbol{\theta}$.

In machine learning classification methods, the vector of parameters $\boldsymbol{\theta}$ are fit to the available training data, in a manner that minimizes the total loss function across the training data samples.

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{n=1}^{N} l(g(\boldsymbol{x}_n, \boldsymbol{\theta}), y_n) \tag{26}$$

Where:

- $\boldsymbol{\theta}$ is a vector of parameters
- N is the total number of samples in the training set
- $\boldsymbol{x}_n$ is a vector of features describing the n^{th} sample in the training set
- y_n is the correct label for the n^{th} sample in the training set
- $g(\boldsymbol{x}, \boldsymbol{\theta})$ is the label assignment function as shown in Equation (24)
- $l(\hat{y}, y)$ is the loss function for predicted label $\hat{y}$ and correct label y

Binary classification problems are a special case where only two possible labels are available for each value of y. These values are generally represented as 1 (true) and 0 (false), with the selection of which value constitutes "true" being somewhat arbitrary. In this case, only a single probability P(y=1) needs to be represented, since P(y=0) = 1 – P(y-1). This simplifies the equations to:

$$g(\mathbf{x}) = \begin{cases} 1 & \text{if} f(\mathbf{x}, \theta) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = P(\, y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}\,)$$

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{n=1}^{N} l(g(\boldsymbol{x}_n, \boldsymbol{\theta}), y_n)$$

(27)

Where:

- $\boldsymbol{\theta}$ is a vector of parameters
- N is the total number of samples in the training set
- $\boldsymbol{x}$ is a vector of features, with $\boldsymbol{x}_n$ describing the n^{th} sample in the training set
- y_n is the correct label for the n^{th} sample in the training set
- $g(\boldsymbol{x}, \boldsymbol{\theta})$ is the label assignment function as shown in Equation (24)
- $l(\hat{y}, y)$ is the loss function for predicted label $\hat{y}$ and correct label y, used to fit $\boldsymbol{\theta}$ to the training data

In the definition of $g(\boldsymbol{x}, \boldsymbol{\theta})$ above, a fixed value of 0.5 is used as the cutoff for the probability of y=1 at which the value of 1 would be assigned. Other decision thresholds can be used, and tuning this decision threshold for specific applications is common practice. Lower thresholds will result in more false positives but fewer false negatives. Higher thresholds will result in more false negatives but fewer false positives.

Machine learning failure prediction problems are a special case of binary machine learning classification problems. In these problems, the y=1 label is generally assigned to a failure event. This further simplifies the definition of f(x, θ) to:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = PoF(\boldsymbol{x}, \boldsymbol{\theta})$$

(28)

Where:

- PoF($\boldsymbol{x}$, $\boldsymbol{\theta}$) is an estimate of the Probability of Failure for sample x, under parameters θ

Where failure prediction relates to a specific period, as it does in pipe failure prediction, $\boldsymbol{x}$ generally includes a time period of interest t, in a manner similar to time series forecasting. This time variable can either be included in x or noted separately. For example, $PoF(t \mid \boldsymbol{x}, \boldsymbol{\theta})$ represents the Probability of Failure during time interval t, for a pipe described by feature vector $\boldsymbol{x}$ and using parameter values $\boldsymbol{\theta}$. This is equivalent to $PoF(\boldsymbol{x}, \boldsymbol{\theta})$ where t is included in $\boldsymbol{x}$. For failure prediction problems, each member of the population generally includes such a period of interest, making the final formulation of the machine learning estimator function:

$$f(t, \boldsymbol{x}, \boldsymbol{\theta}) = PoF(t \mid \boldsymbol{x}, \boldsymbol{\theta})$$

(29)

Where:

- $\boldsymbol{\theta}$ is a vector of parameters
- $\boldsymbol{x}$ is a vector of features
- t is a time period of interest
- $PoF\ (t \mid \boldsymbol{x}, \boldsymbol{\theta})$ is the estimated Probability of Failure during time period t

2.3.1.2 Machine Learning Classification Algorithms

Machine learning classification algorithms can be grouped into six broad categories: Linear Models, Tree-Based Models, Ensemble Methods, Nearest Neighbor Methods, and Neural Networks.

2.3.1.2.1 Linear Models:

Linear models aim to find a linear boundary that separates different classes. Some widely-used linear models include:

- Logistic Regression: Introduced in Cox (1958), this algorithm models the probability of a binary outcome.

- Support Vector Machines (SVM): SVMs transform the input vector x to a higher-dimensional space with the aim of making the classes linearly separable and then finding the hyperplane that maximizes the margin between classes (Vapnik, 1995).

2.3.1.2.2 Tree-Based Models:

Tree-based models create decision trees to make classification decisions. Key algorithms include:

- Decision Trees: Induced decision trees, as described in Hunt et al. (1966), recursively split data based on feature values. Each split is made to maximize some property in the resulting child nodes.
- Random Forest: As described in (Breiman, 2001), a random forest combines multiple decision trees to improve generalizability and reduce overfitting.

2.3.1.2.3 Nearest Neighbor Methods

These methods classify data points based on the class of their nearest neighbors. Notable methods include:

- k-Nearest Neighbors: This algorithm, as described in Fix & Hodges (1989), assigns a class based on the majority class among the k nearest neighbors within a feature space defined by a vector of numerical features.

2.3.1.2.4 Ensemble Methods

Ensemble methods combine multiple base classifiers to improve classification performance. Important ensemble techniques are:

- AdaBoost: Proposed in Freund & Schapire (1997), this algorithm combines multiple weak learners to create a strong classifier.
- Gradient Boosted Trees: This algorithm, described in Friedman (2001), builds an ensemble of decision trees sequentially to minimize a loss function.

2.3.1.2.5 Artificial Neural Networks:

An artificial neural network (ANN) is a network of nodes that each mimic the behavior of a neuron. Each node accepts numerical inputs, multiplies each by a weight, and sums the resulting values. The

sum is then passed through an activation function, and the result is the node's output. The weights for the nodes are the model parameters. The backpropagation algorithm allows the cost function to be minimized by determining the gradient of each weight in the model with respect to the cost function.

The networks are generally arranged in layers, with nodes in a given layer accepting input from the previous layer and sending their output to the next layer. The first layer, known as the input layer, takes its input directly from the feature vector. The final layer, known as the output layer, will have one node per class, with each node's output being the relative likelihood that the item belongs to that class. As with other classification algorithms, the system assigns a predicted label based on the class which received the highest likelihood estimate. Types of ANN considered for use in this project include:

- **Perceptrons**. The original ANN, consisting of a single node using the sigmoid activation function.
- **Multi Layer Perceptrons (MLPs)**. A unidirectional network including one or more fully connected layers (known as hidden layers) between the inputs and the output layers.
- **Extreme Learning Machine (ELM)**. An ANN with (typically) one hidden layer where the input weights to the hidden layer are randomly assigned to speed the learning process and reduce the chances of overfitting (Khozani et al., 2017).
- **Recurrent Neural Networks (RNNs)**. Includes a mechanism for cycles in the network, feeding information back to earlier layers. Models such as Long Short Term Memory (LSTM) and Gated Recurrent Units (GRUs) are used when the population forms a sequence of arbitrary length, allowing for information about past elements to be "remembered" by the system.

2.3.1.2.6 Deep Learning

Deep learning refers to ANNs in which many layers are utilized, forming a deep network. Deep learning concepts considered for this project include:

- **Convolutional Neural Networks (CNNs)**. Introduced by LeCun et al. (1998), CNNs are used when features have spatial relationships. Each node is connected only to a small number of nodes adjacent to its location in the previous layer. All nodes in a given layer

share the same weights. This arrangement results in each layer being a convolution of the previous layer. The combination of weight sharing and limited connections reduces the computational cost of training deeper networks.

- **Embeddings**. An embedding maps a vector onto a smaller vector, with each element in the smaller vector being a function of elements in the larger. Embeddings reduce sparse high-dimensional feature vectors to denser vectors with lower dimensionality. The parameters for the mapping are learned as part of the process of fitting a model to the training data.
- **Attention**. The attention mechanism allows a node to "focus" on the most relevant input. The mechanism uses a set of context variables to generate attention weights. A subset of the node inputs are multiplied by each attention weight, adding more weight to certain input elements. The parameters used to map the context variables to the attention weights are learned as part of the training process.

2.3.1.3 Machine Learning for Pipe Failure Prediction

The objective of this project is to provide quantitative estimates of pipeline conditions using available data. This section provides detailed descriptions of the specific problem formulations to be considered.

2.3.1.3.1 Defining the Population

In pipeline diagnostics, there are several reasonable options for what constitutes a member of the population to be diagnosed, offering varying levels of granularity. We discuss three levels of granularity here, which will be referred to as *stick*, *segment*, and *cohort*. The terms stick and segment are used as defined in Chapter 2.1.1.1.2.

Some studies examine the risk of failure of individual *sticks* of pipe. This is most common for prestressed concrete cylinder pipe (PCCP). The reason for this is twofold. First, the particular failure mechanisms of PCCP allow each stick to degrade and fail at a rate unconnected to the sticks adjacent to it. Second, PCCP tends to be used for large diameter pipes, leading to both a high cost of replacement and a high consequence of failure. Spot repairs are thus practical.

Most utilities use *segment*-level granularity in their systems of record – specifically their Geographic Information System (GIS) and their Computerized Maintenance Management System. This makes segments a natural level of granularity for defining a member of the population for management decisions. It is the definition used most commonly in literature and is the definition of a member of the population used in this study.

Also considered is a *cohort* – a group of segments expected to behave and degrade similarly – as a member of the population. A cohort will generally consist of one material and diameter, will usually be installed in the same general time and area. For transmission mains and trunk mains, a cohort often means a single pipeline from source to destination. For distribution networks, this definition often means the network of pipelines in a neighbourhood or city block. This level of granularity is often the one at which replacement or rehabilitation decisions are made. Consequently, it may be the most pragmatic, although it results in smaller sample sizes than the others.

2.3.1.3.2 Data Availability

Data available for pipeline diagnosis is quite varied. It typically consists of demographic data, environmental data, structured and unstructured maintenance records, time series data from sensors, quantitative test results, and finally data about the network itself (geospatial, and network topology).

Demographic data is the most consistent and has been the most widely studied. It typically consists of a handful of parameters that can be applied to each member of the population:

- Age (or installation date)
- Pipe material
- Joint type
- Pipe inside diameter
- Pipe wall thickness

Other demographic data is sometimes available, such as the pressure class, installation contractor, pipe manufacturer, or design firm. The demographic data is not always reliable, as the records often date back to the early 20th century. The pipe wall thickness, in particular, is often questionable. These generally describe the minimum thickness permitted by the applicable standard. Manufacturers

often produced thicker pipe to ensure compliance, particularly given the wide manufacturing tolerances common decades ago.

Common sources of environmental data are records and tests describing the soil around the pipeline, and what is on the surface above the pipeline. Soil data generally consists of records of the bedding material (what the pipe is laid on) and the cover material (what is placed over the pipe when it is buried), as well as soil chemistry tests made from test pits. Some soil tests, such as Linear Polarization Resistance, can be the primary measurement made for pipeline diagnosis, as they are a strong predictor of corrosion. Surface cover data generally focuses on roadways and traffic loading, as these are predictors of the stress put on pipelines. These have also been extensively studied and shown to be predictive. Weather and climate data are generally accessible but are often neglected in studies that focus on a single locale (and hence share the same weather and climate).

The time series data generally comes from sensors distributed around the network. Common sensors are flowmeters, pressure sensors, and acoustic sensors, with other types of sensors (temperature, chlorine levels, etc.) less common. In medical diagnosis, there is generally a one-to-one relationship between sensors and patients. For pipeline diagnosis, the sensors tend to be sparsely distributed around the pipe network. This presents challenges in identifying which sensors provide data that is relevant to a given member of the population. The frequency of readings can also vary widely. Sensors located in powered facilities, such as treatment plants and pump stations, often take high frequency readings (every minute, every second, or more). Battery powered devices often take much lower frequency readings (every 15 minutes, or hourly). Acoustic devices often take very high frequency readings (above 1000 Hz) at periodic intervals (often daily). These variations make them challenging to integrate into a single model.

Discrete measurements generally require the pipe surface to be exposed. These are often made opportunistically, such as when a pipeline leak or burst requires excavation for repair. Some utilities run proactive programs of exposing sections of pipe for testing. These can be destructive, in which a "coupon" is taken from the pipe for testing, or non-destructive, where sensors are applied to the pipe in-situ. Common destructive measurements are phenolphthalein dye testing for asbestos cement pipe, and the use of pit gauges to measure corrosion pitting. Non-destructive tests include electromagnetic wall thickness testing for ferrous pipes and ultrasonic wall thickness testing for all pipe types. Less common are continuous tests, such as inline electromagnetic tests, and external acoustic tests which

use acoustic wave propagation velocity to measure average wall thickness over intervals (typically about 100m).

Descriptive data of observations is generally found in maintenance databases. These records are filled in whenever a work crew is called in to act on a pipeline. This often includes a categorization of the event (e.g., a leak, a dirty water complaint) as well as a free-form field (such as "notes" or "comments") where observations are described.

The data that describes the network generally comes in two forms. Geospatial data provides a 2-dimensional description of the physical locations of nodes in the network (isolation valves, fire hydrants, intersections of pipelines, etc.) and the pipelines that connect them. The network topology data describes how these features are connected. For example, two pipelines that pass by (one above the other), may be joined together but sealed off by a closed valve, or may have an open connection.

2.3.1.3.3 Diagnostic Objectives

Several pipeline diagnostic objectives have been studied with machine learning approaches. All share the same objective: helping water companies decide when to rehabilitate or replace a section of a pipeline.

One common objective is to predict the remaining life of a pipeline. This problem is ill formed. No consensus exists as to what defines the end of life for a pipeline. Some standards used by utilities include having a set number (often 1, 2, or 3) leaks (or bursts) from a single "asset" in their GIS (usually the length of pipe between to "nodes" in the network). This leads to long stretches of pipeline "dying" long before their shorter counterparts. Some use financial measures, declaring a pipeline "dead" once the net present value of expected future leaks & bursts exceeds the cost of replacing the pipe. Others set a service level threshold (often leaks per 100 km per year) and aim to replace pipelines once they reach this threshold. Still others set a minimum acceptable remaining wall thickness and aim to replace once this is reached.

While a network cannot be trained to predict a measure for which there is no consensus definition, they can be trained to predict the quantitative metrics that are commonly used by utilities to construct these rules. The specific objectives to be studied will depend on the data received. Candidates are:

Estimating the probability of failure. This is either the probability per segment per year or normalized to failures per 100 km per year. For certain types of pipe, the probability of failure per

year per individual stick of pipe is of interest. This probability can be calculated both for the present and as a projection into the future.

Predicting the time to next failure. This is an alternate formulation of predicting the probability of failure.

Estimating the consequences of a failure. This component of risk management would offer tremendous value in risk management. It requires data on the actual cost of past failures, which is rarely stored.

Predicting whether a leak is already present. Many leaks go undetected for long periods of time. Certain sensors (such as acoustic sensors) are used to detect non-surfacing leaks proactively.

Estimating pipe wall thickness. The variable itself is somewhat ill-defined, as different metrics exist. The average thickness, the percentage of metal loss, the maximum pit depth or minimum remaining wall thickness in each stick of pipe, and the linear average of the circumferential minimum, and the results of a crush strength or burst strength test, are all used. Rather than try to define the "correct" measure, a system could simply try to predict what the results would be of a particular type of test. As with other metrics, this calculation can be done for the present state and as a projection into the future.

Predicting the time until replacement. This is predicting human behavior just as much as predicting pipeline behavior. It is analogous to predicting length of patient stay. Interestingly, if it is trained on the actual behavior of leading utilities, it could offer an embodiment of industry best practice.

2.3.2 Prior Work on Machine Learning for Pipe Failure Prediction

Numerous attempts have been made at applying machine learning to pipeline diagnostics. Many of these have been undertaken by private for-profit companies, such as Fracta and Voda, who generally hold their approaches and results as confidential trade secrets and present only highly obfuscated approaches and results (Fracta, 2022). A selection of published results is shown below.

A literature review by Delnaz et al. (2023) provided a comprehensive analysis of asset management analytics for urban water mains, with a focus on the application of machine learning to water pipe break prediction. This paper categorizes various methods used for predicting water main failures and evaluates their effectiveness. It also explores the challenges and future directions in the field,

highlighting the importance of integrating machine learning models with traditional techniques to improve the accuracy and reliability of water pipe break predictions. This study provides a taxonomy for classifying different data-driven models, as shown in Figure 4, grouping them into Deterministic, Probabilistic, and Learning-based approaches.

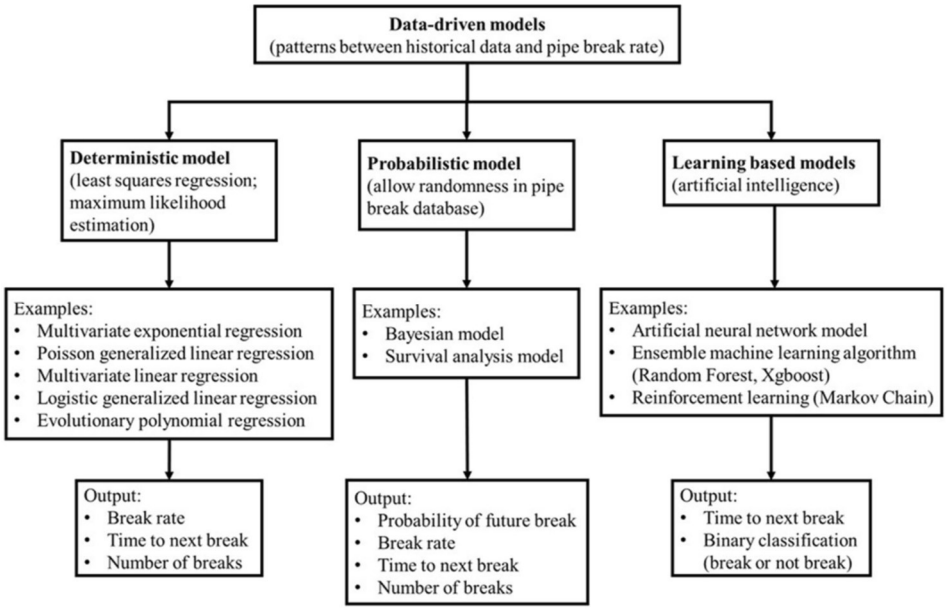


Figure 4: Classification of water main predictive models (Delnaz et al., 2023)

2.3.2.1 Summaries of Prior Studies

Farmani et al. attempted to predict the number of failures among cohorts of pipe from a water distribution system in the UK. They employed a 2-stage cohort formation approach, first grouping pipes by diameter, age, and soil type, and then applying K-Means to cluster these into larger cohorts (Farmani et al., 2017). They then applied Evolutionary Polynomial Regression to predict the number of failures within each cohort. The study obtained RMSE of between 5.5 and 7.5, and R^2 values of between 0.9 and 0.92, depending on the number of clusters (cohorts) used (between 1 and 7).

Wilson et al. conducted a comprehensive review of water pipe failure prediction models, emphasizing the applicability of these models to large-diameter mains (Wilson et al., 2017). They explored various machine learning techniques, examining their effectiveness in predicting pipe failures in complex water distribution networks.

Zhang et al. employed a Bayesian nonparametric machine learning model to incorporate expert knowledge to pipe and failure data from three regions (likely from Australia, as the authors all have .au email addresses). They focused only on "Critical Water Mains" (those of 300mm diameter or greater), giving them a dataset which spanned a 12-year period and included a total of roughly 11,000 pipes and 1,500 breaks. The authors employed a time-based split of training and test data, training on the first 11 years and testing on the 12th. The model achieved AUC scores of either 0.827 or 0.669 (Zhang et al., 2018); both scores are shown in the same cell of the table, with no clear explanation as to what each represents, however a reasonable guess is that these are scores from the train and test set respectively.

Winkler et al. focused on modeling pipe failure in water distribution networks using boosted decision trees (Winkler et al., 2018). Their approach enhanced the prediction accuracy of pipe failures by effectively incorporating complex data relationships. Their available data came from a single midsized with about 40,000 pipes and approximately 20 years of failure history. Their model aimed to predict whether a given pipe experienced at least one failure at any point in time during the study period. The data was split randomly and equally between train and test sets. Imbalanced datasets were handled by random under sampling. Performance was measured using both confusion matrices and ROC curves, with the decision-tree based models achieving AUC scores of 0.9 to 0.93.

A series of studies by Snider and McBean offer significant contributions to the application of machine learning in predicting pipe failures. Their first published study on the topic compared a gradient-boosting algorithm model, an ANN model, and a Random Forest algorithm for predicting the time to the next break of ductile iron pipes (Snider & McBean, 2018). The gradient-boosting model, which combines multiple learning algorithms, was found to outperform the others due to its ensemble nature. They later addressed the issue of data censoring in machine learning models for water main failures, suggesting that survival machine learning models, like the Random Survival Forest, could effectively incorporate censored data and model complex relationships between input and output variables (Snider & McBean, 2020a). They further investigated the relationship between training data size and model performance across several prediction algorithms (Snider & McBean, 2020b). The study considered the task of correctly sequencing a series of pipes in terms of their time to next burst. It compared the performance on different sizes of data sets for XGBoost and simpler methods such as Weibull Proportional Hazard, age ranking, and ranking based on the number of previous breaks. This study found that XGBoost outperformed the simpler methods for large data sets (>5,000 pipe

segments) when at least 10 years of historical data is available and for mid-sized data sets (1,000 to 5,000 pipe segments) when at least 25 years of data is present. In an extension of their work (Snider & McBean, 2021), they explored the potential impact of random survival forests combining the strengths of machine learning with survival statistics. This paper included a case study suggesting a reduction in pipe replacement and repair costs by 14% over the next 50 years was possible using this technique.

Weeraddana et al. utilized machine learning to predict water main breaks in Melbourne, Australia. This study aimed to use a small number of pipeline failure predictors to estimate likelihood of failure in the upcoming 20-year period. The study used a two-step approach, first applying a failure probability estimator based on random forest classification. The classification algorithm was trained to predict failure in a one-year period, matching the recommended PoF(t | x) formulation. It was trained and tested four times, testing on each of the last four years in the data set and in each case training on all data prior to that year. The study observed that failure rates increased during the first 40 years after installation in a relatively linear manner and applied linear regression to project the relative changes in the probability of failure forward in time. The study noted that by selecting the highest 10% likelihood of failure pipes, more than 30% of the future breaks could be identified (Weeraddana et al., 2018). In a follow-up study using the same data, Weeraddana et al. (2019) employed factor analysis to investigate which parameters are predictive, and random forest regression to estimate probability of failure. While the included charts of performance are promising, no quantitative analysis of results is provided, and the paper indicates that 75% of the validation data was also used in the training set.

Almheiri et al. used ANN, ridge regression, and a boosted decision tree to predict time to failure for 1,118 water mains in the Municipality of Saint Foy, Canada over a 15-year period. The study reported R values of 0.84 for the ANN, 0.88 for boosted decision trees, and 0.90 for linear regression (Almheiri et al., 2020). While the paper did not clearly state what the target variable was, from the dates shown it appears to be pipe age at first failure, and that the dataset included only pipes experiencing a failure during the study period.

Al-Ali et al. used a Logistic Regression model aimed at identifying the most critical parameters for predicting the probability of water main failure (Al-Ali et al., 2020). Their data set included nearly 30,000 pipes and monitored for a six-year period, with a total of 1,053 breaks affecting 847 pipe

segments. Their classifier aimed to predict whether any breaks occurred during the study period. Their most significant contribution was the selection of logistic regression, which offers calibrated probability estimates by default, and using these as estimates as the probability of failure for failure risk ranking purposes. Model performance was tested on the training data (no train/test separation was performed), and only the sensitivity (True Positives / (True Positives + False Negatives) was reported, at 74.3%.

Fan et al. investigated the impact of including a range of factors in failure prediction, including the novel inclusion of socio-economic factors, using a database of over 51,000 pipes from the City of Cleveland, USA (Fan et al., 2022). The authors found that gradient boosted trees offered the best performance, achieving an AUC of 0.90 when no time separation was enforced between the train and test sets. The authors note that a time-based separation was also tested, but did not include the results.

Dawood et al. conducted two systematic literature reviews, one focused on water pipe failure prediction and risk models (Dawood et al., 2020b), and the other considering a range of analytical approaches such as machine learning and fuzzy inference for a range of water network management activities. They suggested considering a range of factors such as soil type, traffic loads, construction methods, and contractor experience for improved prediction of pipe failures (Dawood et al., 2020a). They further suggested combining multiple inference techniques into a joint assessment of pipe integrity.

Giraldo-González and Rodríguez studied the prediction of pipeline failures in Colombia's water distribution network, which included just over 60,000 pipes spanning over 1,800 km, and used a study period of seven years (Giraldo-González & Rodríguez, 2020). The dataset used is somewhat suspect, presenting a total of only 69 breaks, reflecting an average of 0.55 breaks / 100 km / year. They considered both the problem of predicting the number of failures in cohorts using three regression models (Linear Regression, Poisson Regression, Evolutionary Polynomial Regression), and also classification of individual pipes' failure status using four machine learning classification models (ANN, Bayes, SVM, GBT). All analysis considered total failures during the seven-year study period. A random splitting of train and test sets was performed, with no time-based separation to confirm the ability of the models to extrapolate forward in time. The study incorporated physical factors (age, diameter, length), environmental factors (moisture content, soil contraction), and operational factors (valve, hydrant, previous failure) as predictors. Top performance on the

regression problem was achieved with the Poisson Regression model, which achieved an R^2 of 0.927 and RMSE of 22.09 on the test data. Classifier performance was measured using confusion matrices, accuracy, F1 Score, and ROC AUC. The top classifier was reported as the gradient boosted tree, with AUC of over 0.99. This extremely high performance is suspect and suggests the possibility of information leakage from the target variable into the features. The paper notes that an important feature in predicting whether a failure occurred was the Previous Failures. Considering that the target variable was whether a failure occurred at any point in the available records, it is plausible that the number of previous failures feature was in fact the number of failures occurring during the study period. This possibility is supported by the fact that these results were inadvertently replicated during our own study, when the number of failures during the target period was inadvertently included as a feature, which also in AUC scores of over 0.99 for a gradient boosted tree classifier.

Khan et al. provided a practical case study using data from 3,779 miles of cast iron pipe in New Jersey, USA. This study worked by grouping pipes into geographic districts and used a rules-based approach to aggregate various failure risk measures for pipes within that district. The model results were used in practice for making water main replacement decisions, with 32% of the highest-risk cohort replaced in the two-year study period. This provides fascinating insight into the practical application of a model, and confirmed that the relative failure rate, as measured by $PoF(t \mid x, \theta)$, is indeed a critical output for practical use. While no standard quantitative measure of performance was provided, a chart showing the actual rate of breaks per 100 km per year among the five risk groupings (Low to High scale) showed a clear progression of increased failure rates in the high-risk groups. The paper further went on to elaborate the cost savings resulting from the work. They estimated 54 breaks per year avoided, at an average cost of $4,000 each, for $216,000 per year of operational costs avoided (Kahn et al., 2020). The utility is permitted to spend $8 in capital costs for every $1 in operational costs saved, which they state is standard practice, thereby freeing up an additional $1,728,000 in capital costs. At their standard rates of $225/foot, this would allow an additional 1.45 miles of pipe replacement per year.

In one of two significant publications emerging from the thesis of Alicia Velasco Robles (2022), Robles-Velasco et al. used Logistic Regression and Support Vector Classification to predict pipe failures, focusing on all pipes in the network rather than just those that had previously experienced breaks. The study was conducted on data from the City of Seville, with 3,800 km of pipe, with a study period of seven years including 4,393 breaks (Robles-Velasco et al., 2020). They focused on

binary classification to determine whether a pipe would break or not. They used the model's label assignment scoring, ranging between 0 and 1, to represent the probability of failure, to facilitate use of this probability of failure estimate in other common pipeline management techniques. Time-dependent separation of the training and test set was employed, with the first five years used to train and the final two years used to test. The choice of including differing periods (five years vs two years) in the training and test set is of questionable validity, as it means the model is trained to estimate the probability of failure in a five-year period and tested on its ability to estimate the probability of failure in a two-year period. It is a reasonable assumption that these will have an almost perfect rank correlation, so metrics such as the AUC should be unaffected. Under-sampling was utilized to balance the dataset, which was found to substantially improve model results. This study noted that replacing diameter with its logarithm substantially improved the performance of both of these methods, improving AUC performance from 0.863 to 0.876 for logistic regression, and from 0.798 to 0.855 for Support Vector Classification; this suggests a consistent but non-linear relationship between diameter and risk of failure, which is further explored in Chapter 4.2.1 and discussed in Chapter 6.1.2. Another interesting advance presented in this paper was tuning the decision threshold to measure performance of the classifier when it was permitted to select only a small percentage of the network for replacement. This approach can be generalized to the Cumulative Lift performance metric at a low percentage, as described in Chapter 3.6.3 .

In their follow-on study using the same data, the authors explore predicting failure in each of three successive one-year periods using a classifier chain (Robles-Velasco et al., 2023). This explicit formulation of estimating the probability of failure within a given time period PoF(t | x) matches the approach described in Chapter 3.5 and is selected for this study. The study further explores the value of selecting the highest-risk pipes for replacement, providing estimates of the percentage of failures that could be avoided by replacing the 5% (32.1% of failures avoided), 10% (51.4% of failures avoided) or 15% (54.0% of failures avoided) of the pipes respectively.

Aslani et al. employed spatiotemporal data to predict water main failures using main break records from 2015 to 2020 in Tampa, Florida, with over 3,000 failures recorded. They used machine learning models to predict water pipeline breaks with inputs of spatiotemporal data. Prior to applying the machine learning models, they conducted spatial clustering to identify the relative rates of pipe failures in geographic regions (a "hotspot level") and provided this as a feature for use in the machine learning model (Aslani et al., 2021). While the "hotspot level" is a novel concept, it should be noted

that this implementation included data from the test set in defining the hotspot level, making this an example of "information leakage" from the target variable into the features. Furthermore, the train and test data were separated randomly, with both datasets spanning the same period. As a target variable, they chose the average number of failures per month during the six-year study period. This regression model calculates the expected number of failures per month, allowing for easy aggregation of results into groups or cohorts of pipe. The study reported Root Mean Squared Error (RMSE) in the number of failures per pipe per month in a six-year period (again, a unique and unusual target variable) ranging from 0.083 for an ANN down to 0.007 for a boosted regression tree estimator. They also measured performance by ranking pipes based on their expected number of breaks and providing the percentage of breaks captured by selecting a given percentage of pipe length. The charts presented show that by selecting 10% of the pipe length, approximately 22% of breaks could be captured. Furthermore, they presented an aggregate "Area Under the Break Capture Curve" as a novel performance metric, which is calculated in a manner analogous to area under the receiver operating characteristic curve but uses percentage of length for the x-axis.

The thesis work of Mohammad Amini explored the application of machine learning models to predict failure status in data from 13 Canadian utilities (Amini, 2021). Extensive manual data cleansing was performed, including manually mapping the categorical variables used by each utility to a common standard. The models aimed to classify each pipe based on whether a break occurred at some point during the study period, which varied by utility. Note that an "age" feature was included, which was either the age at time of first recorded break or the age as of the last year in which monitoring occurred, which has the potential for information leakage from the target variable into the features. Top performance was reported with XGBoost, achieving F1 scores of between 0.07 and 0.88 across the 13 utilities, with each utility's data being used to train and test separately, but no time separation between train and test data. A final test was reported on cast iron pipe only, training a single model on data from all utilities. This model achieved an F1 score of 0.72, with the authors reporting no change in the overall (aggregate) F1 score, but an improvement in the F1 score for each utility, with the average improvement well over 10%.

Amini et al. compared random forest, logistic regression, and decision tree models on data from Saskatoon and Waterloo, Canada (Amini & Dziedzic, 2022). Extensive data cleansing was applied, considering only pipes with length less than 200m, ages under 80 years, and materials from the five most common materials in these networks (AC, CONC, DI, PVC, and PE). Certain unusual statistics

were reported from the data, such as mean values greater than the max value for diameter, length, and age in one of the contributing utilities. A random train / test separation was employed on the roughly 35,000 pipes and 19 years of records, with no time-based separation of train and test data enforced. The study notes that 12 binary variables were created as to whether a failure occurred during each month of the year, however the choice of target variable is not explicitly stated. It may have been whether a given pipe would fail during the entire study period, whether a given pipe failed in any instance of a given month within the study period (i.e., any January failures, any February failures, etc.), or whether each month in the study period experienced a failure (i.e., any failures in January of 2000, any failures in February of 2000, etc.).

Karimian et al. attempted to predict the number of breaks per pipe over the course of a >50-year study period using a variety of regression models. The dataset consisted of water infrastructure in Montreal, including 5,045 km of water linear assets with a total of 22,735 pipe breaks recorded between 1972 through to an unspecified end date (assumed to be shortly before 2020). The models aimed to predict the number of breaks for each pipe segment, utilizing variables such as pipe length, diameter, age, and material (Karimian et al., 2021). While the paper refers to an "annual" number of breaks, it is unclear whether this is the average break rate per year or if each pipe-year combination is predicted separately. The top performing model was evolutionary polynomial regression (EPR), with the best-performing model reaching an R^2 value of 89.35%, with the most plausible interpretation being that the target variable was the number of breaks per segment in total during the >50-year study period.

Zakikhani et al. employed multiple linear regression for failure prediction on 20 years of records for gas transmission pipelines in the USA. It employed multiple known predictive values from physical tests (e.g., operating temperatures, hoop stress, soil moisture content, size of cathodic protection anodes, redox potential, etc.), to predict the observed values of their continuous dependent variable y (Zakikhani et al., 2021). It is never clearly stated what y represents (time to failure, consequence of failure, and failure rates are all mentioned), or how members of the population are defined (these are less clear on long-distance gas transmission pipelines), however the authors report R^2 values of 0.93 and 0.75 in two different climate regions.

While sewer pipes are easier to access than water pipes and commonly inspected using closed circuit television, Mohammadi et al. used K-Nearest Neighbors to predict the condition of

uninspected sewer pipes. The study considered over 30,000 pipes in the City of Tampa, with a single pipe being clearly defined as a stretch from one manhole to another. Many of the same features (age, material, diameter, length) were employed as are commonly used in water pipe failure prediction. Sewer pipes benefit from a standard 1 through 5 condition classification scale, and the model was trained to distinguish between good (condition 1, 2, or 3) and poor (condition 4 or 5) pipes (Mohammadi et al., 2021). While no time split was enforced between training and test sets, extrapolation forward in time is not as significant for this particular application, which aims to assess the current condition of uninspected pipes rather than to predict future failures. Performance was measured by area under the ROC curve, with an AUC of 0.89 achieved, with age and length being the most important features.

Published results from a for-profit model (Fracta, 2022), while not disclosing their approach, did include a Cumulative Lift chart plotting portion of pipes vs portion of breaks. Manual examination of this chart shows a Lift at 10% of roughly 4x, which is one of the key metrics selected for this study.

Barton et al. attempted to improve performance of a pipe break classification model by grouping pipes together into 1km intervals of pipe with the same material, diameter grouping, and age grouping. The study considered 14 years of failure history on approximately 40,000 km of pipe in a UK city, considering only the four most common types of pipe and grouped steel and ductile iron together as a single material. The study noted that they were able to create over 80,000 cohorts from the 40,000 km of pipe, and that the actual average cohort length was less than 500m, suggesting that more complex rules were implemented to define cohorts than were explicitly stated. The study did not state whether they aimed to predict whether a failure occurred on a per-unit-time basis or over the entire duration of the study. A random train/test split was used, with stratified random sampling to ensure balance of materials across the train and test set. Employing a gradient boosted tree, they were able to achieve an AUC of 0.89. In a novel contribution, performance in estimating the actual probability of failure was also provided using the Brier's Score (essentially the mean squared error in the probability estimates), and reached a score of 0.007 (Barton et al., 2022b). The strong Brier's Score, combined with the fact that gradient boosted trees do not naturally produce calibrated probability estimates, suggests that a method of calibrating the probabilities was applied; however, this method was not discussed in the paper.

Beig Zali et al. attempted to predict failures on a data set from a UK utility with nearly 400,000 pipes and over 18,000 failures over 26 months (Beig Zali et al., 2024). The authors considered only cast iron and asbestos cement pipes for the study, creating a more homogeneous population of the materials deemed to be most at-risk in that network. Segmentation was performed first manually on material, length, diameter, and age, by examining charts of each of these continuous variables versus the cumulative numbers of breaks. Small clusters were grouped together, and large clusters were further subdivided using of k-means. It should be noted that this is an instance of information leakage from the target variable into the features, as both the training and test data was used to create the clusters. Classification was then performed using logistic regression, random forest, and gradient boosted trees. It is not clear whether the study trained a separate classifier for each cluster to predict failure of each pipe segment, or if they trained a single classifier to predict whether a failure occurred within the cluster. The authors compared various methods of class rebalancing and reported the highest AUC scores of 0.803 with the use of undersampling and the gradient boosted tree model.

Omar et al. employed a range of machine learning classifiers to predict breaks for the City of Kitchener, Canada, on roughly 15,000 pipes over a period from January 1985 to January 2021. This study integrated expert knowledge with the machine learning approach via the inclusion of a "Condition Score" variable incorporating factors such as material and history of breaks (Omar et al., 2023). Note that there appears to be information leakage from the target variable into this feature: the target variable was whether any break occurred during the study period, and the number of recorded breaks contributed 50% of the weighting to the Condition Score feature. No time separation was enforced between the test and training sets. The top performing model was the random forest approach, which achieved an Area Under the Curve of 0.814 on all pipe materials, and of 0.862 when a model was trained and tested on cast iron pipe only.

Fan et al. employed a Long Term Short Memory (LSTM) time series model to predict the total number of breaks per month across an entire water network, using 35 years of data from Cuyahoga County, USA (Fan et al., 2023). They found that the LSTM approach yielded slightly lower bias (13.3% lower) and narrower confidence intervals (44.7% smaller) than the ARIMAX method used for comparison.

Finally, two recent studies have begun to explore the ability of machine learning models to extrapolate from one utility to another. Both include elements of the approach used in this study, which requires extrapolation both forward in time and to a new location concurrently.

Chen et al. conducted a study using data from six water utilities in the USA, all contributing break data from the period of 2005 to 2018 (Chen et al., 2022). Each utility had between 36,000 and 265,000 pipes, and between 500 and 22,000 breaks. Strict time separation was enforced between the training and test sets, with data from 2005 to 2015 used for training, and from 2016 to 2018 for testing. As a target variable, this study elected to train based on whether there was at least one break recorded on that asset or any of its neighbors within 60.96m (200 feet), during the following three-year period, but to test using whether there was at least 1 break recorded on that asset itself during a three-year period. In addition to pipe and environmental features, both time lag (number of breaks in prior 1, 2, 3, 4, and 5 years) and spatiotemporal lag (breaks on the asset or its neighbors with 60.96m in the prior 1, 2, 3, 4, and 5 years) were used. Several data inclusion cases were tested and compared: using each utility's data in isolation (either all of it, or the most recent five years only), supplementing it with all data from the other five utilities inclusively, or supplementing it with an equal amount of data from other utilities in a manner matching the pipe material distribution of that utility. Area under the curve for break capture was selected as the performance metric, with the models tested being random forest and two different implementations of gradient boosted trees. Performance among the six utilities ranged from AUC of 0.73 to 0.90 (average 0.81) using random forest, and from AUC of 0.54 to 0.90 (average 0.79) using boosted trees. Limiting the training data to the most recent five years degraded performance across all algorithms, and particularly on the boosted trees. Supplementing with other utilities' data, however, provided mixed results, with no overall improvement in performance measured. The authors concluded that data quality was more significant than data quantity in training a model.

A promising study Daulat et al. (2024) explored the potential of machine learning models to predict time to failure for pipes in water distribution systems beyond the systems whose data was used to train the models. It specifically investigated whether smaller utilities can leverage data and models from larger or multiple utilities to enhance their predictive capabilities. The study utilized datasets from nine Norwegian utilities to train both multi-utility ("global") and single-utility ("local") models using Random Survival Forest. It examined the generalizability of global models and the transferability of local models in predicting pipe failures of utilities not included in their training

datasets. A single set of data was used for both training and testing purposes, with extensive data cleansing removing 108,000 out of the total 139,000 pipes (22% of the population) via numerous exclusions (e.g., pipes installed before 1945, pipes less than 1m in length, pipes with anomalous installation dates). To fit the survival curves, the authors defined a failure state as a single break and registered a "new" pipe (with more previous breaks) each time a failure occurred. The findings suggested that global models offer promising accuracy in prediction, with performance measured by the Concordance Index (somewhat analogous to Area Under the Curve for classification models) of 0.82 for the local model and 0.80 for the global model. The study also delved into survival curves and variable importance analysis, providing insights into the factors influencing pipe deterioration and the feasibility of model sharing among utilities. The authors noted that "It would be interesting to check the performance of the global models for utilities in other countries with similar and different climate, geography, and historical network evolution to test the limits of such a transferability," (Daulat et al., 2024) which is undertaken in our study, as described in Chapters 3 through 6.

2.3.2.2 Comparative Analysis of Prior Work

These numerous publications over since 2017 in the application of machine learning to water pipe failure prediction have shown several consistent findings, as well as certain limitations.

Decision tree-based models have consistently proven effective on this problem. A wide range of types of machine learning models have been applied in the various studies. Where direct comparisons have been made between different models, the most frequent top performers have been random forest and gradient boosted tree models. These belong to the same family of models which utilize multiple learned decision trees to make predictions or classifications. Models in this family share traits of being able to represent non-linear and even non-monotonic relationships between features and the target variable, as well as robustness to certain data quality problems such as outliers in numerical features.

While performance metrics have varied from study to study, a common performance metric used has been area under the curve (AUC). A summary of comparable results from the literature in which AUC was reported is provided in Table 6. Values achieved by these various studies all fall into the range of 0.8 to 0.9 in their AUC scores, with an average value among the reported studies of 0.85. In cases where a time split is applied (i.e. where the training data is drawn from a time period prior to test data), the range narrows to between 0.81 and 0.88.

Table 6: Summary of performance by AUC reported in the literature.

Study	Model	AUC	Target Variable	Time split?
Omar et al. (2023)	Random Forest	0.81	Any failures in a 35-year period.	n
Robles-Velasco et al. (2020)	Logistic Regression	0.88	Failure in a 2-year period	y
	Support Vector Classifier	0.86	Failure in a 2-year period	y
Zhang et al. (2018)	Bayesian nonparametric	0.83	Failure in a 1-year period	y
Winkler et al. (2018)	Gradient boosted trees	0.90	Any failures in a 20-year period	n
Fan et al. (2022)	Gradient boosted trees	0.90	Any failures in records	n
Barton et al. (2022)	Gradient boosted trees	0.89	Failure in a group of pipes in a 14-year period	n
Beig Zali et al. (2024)	Gradient boosted trees	0.80	Failure in a group of pipes in a 26-month period	n
Omar et al. (2023)	Random forest	0.81	Failure in a 35-year period	n
		0.86	Same, on cast iron only	n
Chen at al. (2022)	Random Forest	0.81	Break in a 3-year period; avg. across six utilities	y

Some consistencies have also emerged with respect to the top predictive features. Diameter and material were noted as top predictive features in each study. The features, which are tracked by nearly every utility, are clearly useful in predicting pipe failure with machine learning. Age was often cited as a top predictor, but not in all cases. In some publications there was in fact no correlation observed between age and failure rate.

Limitations of the prior work include data limitations, methodological limitations, and limitations of practicality. The data sets for each study came from a single source (either one utility or at most several utilities in the same country), preventing the testing of transferability of models and findings to other regions. Different data sets were used in each study, which prevents the direct comparison of modeling methods employed in each study. Further limiting the direct comparison of modeling methods across studies is the wide variability in the target variable used (pipe vs cohort, and duration of time in which the presence or absence of a break is to be predicted). Many of these target variables (such as whether or not a break will occur in a 35 year period) are poorly aligned with utility decision making practices, limiting their practical value to model users.

2.4 Chapter Summary

This chapter provided a review of the literature relevant to the application of machine learning for pipe failure prediction. It first provided background and context for the construction and management of water networks and water pipe failures, including clear definitions of what constitutes a pipe and a failure within this study. A review was then provided of the historically established engineering methods of water pipe failure prediction, and how these are used to support water network management decisions. Lastly, an overview of machine learning methods for classification problems was provided, along with a review of their application for pipe failure prediction as reported in the literature.

Chapter 3 describes the methodology applied in this study for the development and testing of a generalized machine learning model for pipe failure prediction. It describes the process of collecting, preparing, and analyzing the data for the study. It then describes in detail the three layer model developed in this study, and the testing process employed to evaluate the practical applicability of the approach.

Chapter 3

Study Methodology

3.1 Introduction to Methodology

The methodology of this research involved the following steps:

- Data collection on pipes and breaks from utilities
- Datra integration into a standard data model
- Exploratory data analysis
- Machine learning model development and training
- Evaluation of the model performance

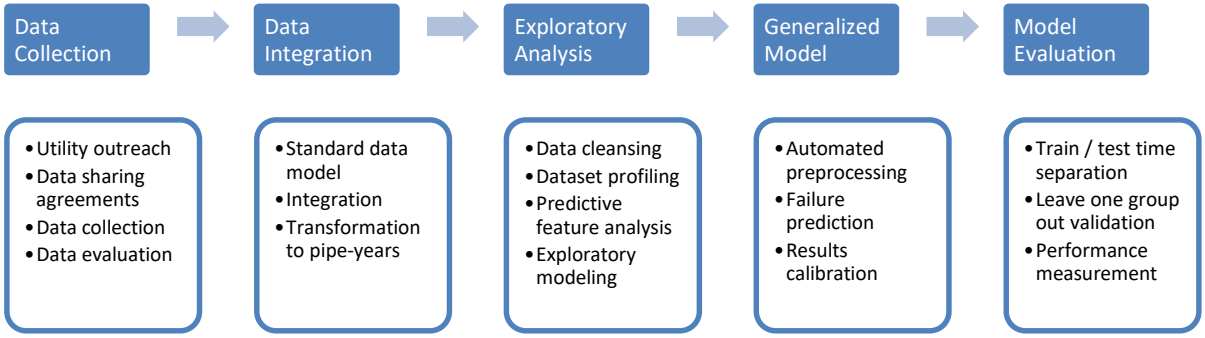


Figure 5: Summary of research project activities.

The data collection stage involved gathering data from multiple water utilities representing a range of geographic and environmental regions. This was an extensive process, taking over two years to complete. Most utilities were reluctant to share data for a variety of reasons. For many it was simply a matter of time and data availability. For other utilities, data security was the primary concern. Data security concerns stemmed from either privacy concerns with respect to the utilities' customers, or security concerns with respect to information about the physical infrastructure. In some cases these concerns were addressed via data sharing agreements. Upon receipt of data, each contribution was evaluated for suitability for the study.

Data integration required bringing the datasets provided in disparate formats into a single common data model. Data was provided in a variety of formats, including .csv files, Excel files, and shape

files. In the case of shape files, a geolocation step was required to match break events to pipes prior to integrating the data into the standard model.

The objective of the exploratory analysis stage was to inform the design of the machine learning model. This involved profiling the data, and examining the relationships between the individual predictive factors and the target variable. Observations from this stage supported decisions regarding feature engineering, normalization, and selection of model types.

Creation of the generalized machine learning model for pipe failure prediction was the primary objective of the study. This involved creation of the proposed model. To ensure repeatability of the study, no manual data cleansing was applied to the data provided by each utility. Minimal preprocessing was performed, and all of it done in an automated manner consistent across all utilities.

Model evaluation was the final stage of this project. The focus of this stage was to ensure that the model extrapolates both forward in time to periods not included in the training data, and also to new utilities which did not contribute training data.

3.2 Data Collection

The best option for data collection is to gather actual pipeline data and failure results from water utilities. Previous research and publications consist of either analysis of data from a single utility, or of collation of survey results from multiple utilities. Survey results have generally been confined to one country. Collecting and preparing a large and geographically diverse data set was a major aim of this research.

Data collection from water utilities was the longest lead time component of this project. Utilities are highly protective of their data, and most participants required a lengthy negotiation process for access to data. Even upon completion of these negotiations, some utilities were unable to collect and share data appropriate for the proposed analysis.

By focusing effort on data collection and preparation, this project has collected usable data from seven utilities spread across four countries in North America, Europe, and Asia. The data contains 174,870 breaks, and 640,226 pipe segments representing 29,626 km of pipe. With a cumulate total of 13,673,295 segment-years of break history, this may be the largest data set used to date for academic purposes.

Data collection was performed in a four-step process of Outreach, Data Sharing Agreements, Data Collection, and Data Evaluation. The number of target organizations at each stage was as shown in Table 7.

Table 7: Data Collection Process Stages

Stage	Number of Organizations
Outreach	27 organizations approached
Data Sharing Agreements	12 organizations responded and opened negotiations
Data Collection	8 organizations provided some form of data
Data Evaluation	7 organizations provided usable data 6 organizations provided data fully usable

3.2.1 Step 1: Outreach

An outreach process was conducted with outreach to a large number of water utilities, research groups, industry groups, and private technical service companies inviting them to participate in the research. Outreach candidates were identified via the existing relationships from members of the research lab. The groups in Table 8 were contacted and invited to participate in the study.

Table 8: Groups Invited to Participate in Study

• Anglian Water	• Brabant Water	• Ontario Clean Water Agency
• American Water	• Suez Environment	• Tacoma Water
• Toronto Water	• Town of Aurora	• Dunea
• The City of Hamilton	• York Region	• KWR [Research Group]
• Vitens	• American Water	• Singapore Public Utilities Board (PUB)
• Waternet	• Halifax Water	• San Diego County Water Authority
• The Sewerage and Water Board of New Orleans	• Washington Suburban Sanitary Commission	• American Water Works Association (AWWA) [Industry Association]

• International Water Association (IWA)	• Virginia Tech WaterID Project [Research Group]	• Center for Advancement of Trenchless Technology [Research Group]
• Peel Region	• City of Winnipeg Water & Waste Department	• City of Kitchener

3.2.2 Step 2: Data Sharing Agreements

Water utilities are generally quite protective of their data. Most participants required Non-Disclosure Agreements or Data Sharing Agreements between the utility and the university prior to sharing data. Negotiation of these agreements took between four and nine months. The groups shown in Table 9 responded to outreach efforts, and engaged in discussions for access to data:

Table 9: Groups Responding to Outreach Efforts

• American Water	• Ontario Clean Water Agency
• Toronto Water	• KWR [Research Group]
• The City of Hamilton	• Singapore Public Utilities Board (PUB)
• Vitens	• Washington Suburban Sanitary Commission
• Waternet	• Virginia Tech WaterID Project [Research Group]
• Dunea	• Peel Region

3.2.3 Step 3: Data Receipt

Upon successful negotiation of data sharing agreements, a standard data sharing request document was shared with each organization. Some organizations were unwilling to share full data for security reasons and opted to aggregate the data prior to sharing. Others were unable to gather and share all of the requested data. Of the 12 groups that engaged in discussions for access to data, the eight shown in Table 10 provided data in some form.

Table 10: Groups Contributing Data to Study

• American Water	• Dunea
• Toronto Water	• Singapore Public Utilities Board (PUB)
• The City of Hamilton	• Waternet

• Vitens	• Peel Region

3.2.4 Step 4: Data Evaluation

Once received, each data set was evaluated for usability and appropriateness for the research project. The following data sets were received from water utilities.

- **Toronto Water: Fully Usable**. A database consisting of a Pipe Segments table and Breaks table, with an existing key variable (LD_ID) linking each break to a pipe. This was immediately appropriate for the proposed analysis.
- **The City of Hamilton: Fully Usable**. Two Excel files, one containing a Pipe Segments table and the other containing a Breaks table, with an existing key variable (COMPKEY) linking each break to a pipe. This was immediately appropriate for the proposed analysis.
- **Dunea: Not Usable**. One Excel file consisting of a Breaks table and aggregate lengths by material, and only for large diameter transmission mains. It was not appropriate for our proposed research, as it did not provide sufficient detail for Segment-Level analysis, nor was the aggregation sufficiently granular for Cohort-Level analysis
- **Waternet: Fully Usable**. One Excel file with a Break table, and a GIS Shape File with a Pipe Segments table, but no key variable linking the two. This was appropriate for the proposed analysis with geospatial preprocessing to identify the closest pipe to each break.
- **Singapore Public Utilities Board: Fully Usable**. Two Excel files, one containing a Pipe Segments table and the other containing a Breaks table, with an existing key variable (WATERMAIN_) linking each break to a pipe. This was immediately appropriate for the proposed analysis.
- **New Jersey American Water: Fully Usable**. A GIS file with a Break table, and a GIS Shape File with a Pipe Segments table, but no key variable linking the two. This was appropriate for the proposed analysis with geospatial preprocessing to identify the closest pipe to each break.
- **Peel Region: Fully Usable**. A database consisting of a Pipe Segments table and Breaks table, with an existing key variable (WM_COMPKEY) linking each break to a pipe. This was immediately appropriate for the proposed analysis.
- **Vitens: Limited Usability.** A single Excel file containing a Breaks table together with the Pipe Segment details for the Pipe Segment associated with the break. While it contained the

required data for our analysis, it included only Pipe Segments that suffered a failure. This selection bias meant that it had to be handled carefully during analysis to ensure that this bias did not skew the models, the descriptive statistics, or the factor analysis. It was appropriate for the proposed analysis in limited usage.

3.3 Data Integration

To train and test predictive models, the data from each utility needed to be integrated by transforming each into a consistent data model and then integrated into a single database. The data received posed several challenges to this process. First, some utilities do not make a practice of associating break records with pipe records. Second, while certain core features were present in all the pipe tables (material, diameter, length, and installation year), there was considerable variety in the additional features included. Third, even when the same information was contained, the column names, units of measure, date formats, acronyms, and languages varied considerably. Fourth, many of the Break tables were generated as a full extract from the utility's maintenance records database, and included both actual pipe break events (Leaks, Bursts, Breaks, etc.) and other maintenance activities (meter reads, service connections, surface flooding, sewer overflows, etc.).

The data model used and integration process applied to transform the various data sets into this common data model are described in this chapter.

3.3.1 Standard Data Model

Prior to receiving utility data, a conceptual data model was laid. This model reflects data which has previously been shared with the researcher during various instances in his professional career. This preliminary design consists of the four tables shown in Figure 6.

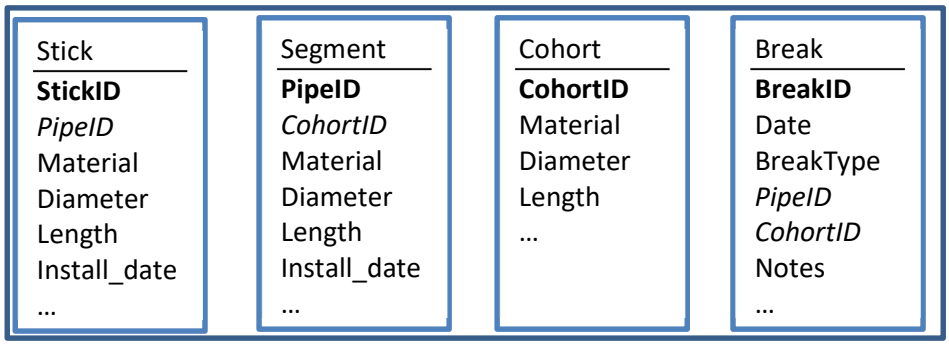


Figure 6: Schema of planned data model. Table headings are above the line, primary keys are in bold, and references to other tables are in italics.

Raw data supplied by utilities was expected to vary. It could come at the Stick (from pipe joint to pipe joint), Segment (one entry in the utility's asset records), or Cohort (a group of pipes expected to degrade similarly over time) level, as described in Chapter 2.1.1.1.2. Data received at the Stick level would require aggregation prior to usage. Data received at the Cohort level would be usable only for testing the trained models' ability to forecast total numbers of breaks in a cohort.

Stick level data is often only available for large diameter pipelines that have been subject to inline inspection (usually PCCP or Steel). In practice, no utilities were willing to share stick level data for this project. The final data model used reflecting the data actually received is shown in Figure 7.

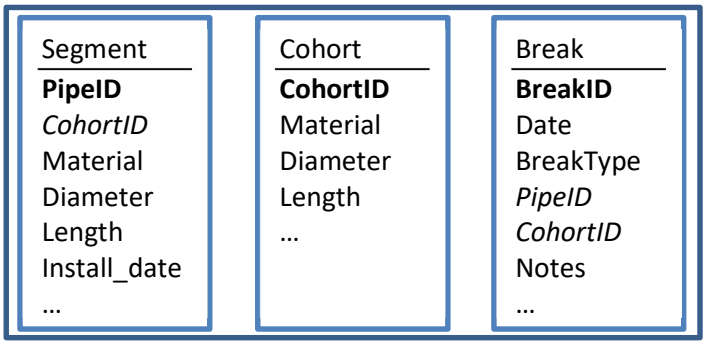


Figure 7: Schema of final standard data model. Table headings are above the line, primary keys are in bold, and references to other tables are in italics.

The level of time granularity for analysis is years. Hence a single member of the population for analysis is a Segment-Year tuple, with Cohort-Year tuples also used for validation.

3.3.2 Data Set Merges

Substantial data integration was required to bring all the data sets into conformance with the proposed data model described in Chapter 3.3.1. Where possible the data integration process was done using scripts or code to ensure repeatability of the exercise. Scripts and code are in the form of python and SQL queries and are executed using the processing software Dataiku. The steps followed for each data set were as follows:

- **Data Integration Step 1**: Geospatial Preprocessing. This step involves linking Break records with Pipe records based on geospatial proximity. It was needed only for Waternet and New Jersey American Water to provide a key variable to link each Break record with a Pipe Segment record.
- **Data Integration Step 2**: Extra raw data into comma separated value (csv) format.
- **Data Integration Step 3**: Import csv files into PostgreSQL tables with standard column names.
- **Data Integration Step 4**: Normalize the units, date formats, and data types.
- **Data Integration Step 5**: Merge Data from all utilities into common pipe and break tables.

During these steps, minimal data cleansing and transformation was undertaken to allow the steps to move forward. Where possible, these operations were deferred to after the end of the data preparation pipeline, to allow them to be performed on the merged data from all utilities. Certain data cleansing operations, such as removing or correcting improper date strings, were required during the integration phase. These steps are described in detail to ensure repeatability of this research.

A data conversion script was developed in Python to generate the Segment-Year tuples for analysis. The number of failures in each tuple has been calculated as the target variable.

3.3.2.1 Data Integration Step 1: Geospatial Preprocessing

Two of the data sources (Waternet and New Jersey American Water) were received in the format of a GIS Shape File for the pipe data, and an Excel file for the breaks data. No linking field was provided to associate each break with a particular pipe.

Geospatial analysis was used to associate breaks with pipes. This stage in the integration process required manual execution of several steps rather than via code or scripts. This was accomplished via a four-step process, as outlined and further detailed below:

1. Convert each address to a latitude and longitude
2. Convert the latitude and longitude to an X and Y coordinate
3. Identify the pipe (a Line object) that is closest to each break (a Point object)
4. Record the pipeid associated with each breakid

After applying these steps, each break record with a valid address was associated with a single pipe via the PipeID field. This enables association of pipe and break records using this field in code and in SQL join operations.

3.3.2.1.1 Convert Each Address to a Latitude and Longitude

The process of converting a location description, such as an address, to a geospatial location, such as latitude and longitude, is known as geocoding. The website geocode.xyz provides batch geocoding services and was used for this conversion. The data was transformed to a format with one column. The column contains the full address, in the format "<House Number> <Street Name>, <City>, <Country>" as illustrated by the example "33 Haarlemmerstraat, AMSTERDAM, Netherlands" (note this is an illustrative example not in the actual dataset).

Many of the address records provided by Waternet showed a house number of 0. This was assumed to be a sentinel value inserted into the Waternet database in cases where no house number was present in the original maintenance notes. The presence of the 0 caused the geolocation service to fail; however, when the numbers were removed, the service provided a geolocation of the lowest numbered house on the street. A manual review of the Waternet GIS file showed that nearly all streets have only a single water main. As such, using the geolocation of the street resulted in a correct mapping of break to pipe in nearly all cases, making it appropriate for this analysis.

This process was able to identify a geolocation for 99.5% of the break locations in the data.

3.3.2.1.2 Convert the Latitude and Longitude to an X and Y Coordinate

This conversion was done using the open-source GIS software package QGIS. Latitude and longitude of the break locations was transformed to X and Y coordinates using the transformation "Inverse of Amersfoort to WGS 84 (4) + RD New." According to the QGIS documentation, this transformation provides accuracy of 1 meter for onshore Netherlands addresses.

3.3.2.1.3 Identify the Pipe (a Line Object) That Is Closest to Each Break (a Point Object)

This matching was done using the QGIS geoprocessing feature. First, two Layers were added to the project: one for the pipes (Shape File provided by utility) and one for the Breaks (the break information, together with the X and Y coordinates created in the previous step). The matching was performed using the command accessible via the "Processing → Processing Toolbox → Vector General → Join attributes by nearest" command.

To minimize the number of false associations, a maximum distance of 100m was permitted between the break location and the pipe object. Note that the break location provided was the address of the nearest building, not the location of the break itself. Given this, the threshold was manually tuned to capture associations between buildings located somewhat distant from the road but to exclude break locations which were recorded in an area not close to any pipe.

3.3.2.1.4 Record the PipeID Associated With Each BreakID

This was performed to enable future association of records via simple join operations.

This process was able to associate a PipeID with 98.0% of those Break records of a type that corresponded to an actual pipe break event, and 67.9% of all the Break records provided. The higher proportion of actual break events is due to many non-break events either having no associated address or an address that is distant from any water pipeline.

3.3.2.2 Data Integration Step 2: Extract csv files

This step involved opening each file provided by the utility in software appropriate for the native file format. The software used included QGIS, Microsoft Access, and Microsoft Excel. Each software included a function to export or save to csv format, which was used to generate these files.

With all data in a standard csv format, all further processing was performed using code (Python) and queries (PostgreSQL). The code and queries were stored and orchestrated using the free version of the Dataiku data processing software. This software visually organizes data tables (shown as squares) and data processing steps (shown as circles) in a flow from left to right. This visual representation of the data integration flow for Steps 3 through 5 on the Pipes is shown in Figure 8.

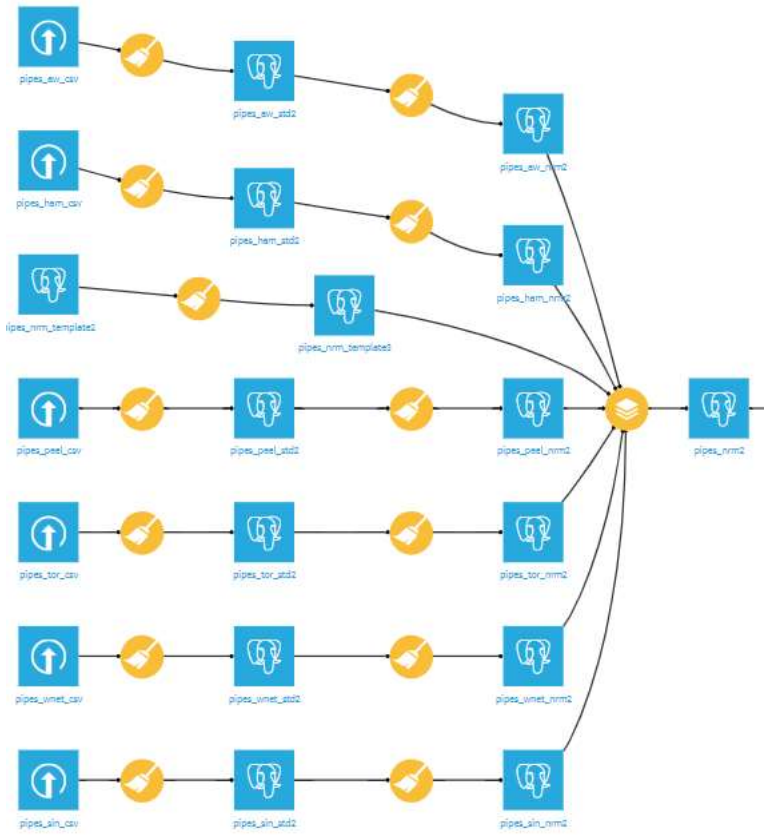


Figure 8: Visual representation of the Data Integration processing steps 3 (standardize), 4 (normalize), and 5 (merge) for the Pipes tables. Squares show tables, and circles show data processing steps.

The corresponding visual representation of the data integration flow for Steps 3 through 5 on the Breaks is shown in Figure 9. This flow is considerably more complex. The greater variety of data structures in which utilities provided breaks data required more processing steps to standardize and normalize.

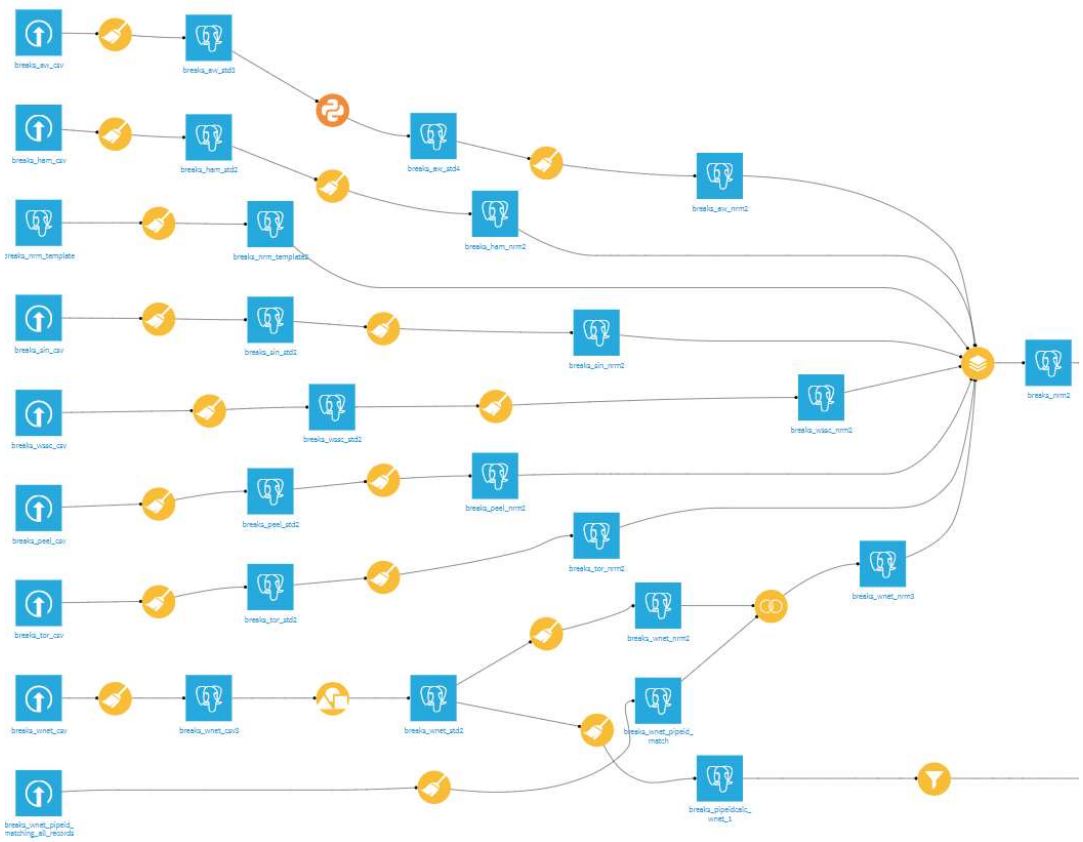


Figure 9: Visual representation of the Data Integration processing steps 3 (standardize), 4 (normalize), and 5 (merge) for the Breaks tables. Squares show tables, and circles show data processing steps.

3.3.2.3 Data Integration Step 3: Import to PostgreSQL Table in Standard Data Model

This step involved converting each csv file into a database table using a standardized naming scheme. Data types were kept as text during this step, with type conversions deferred for Step 4.

The standard data model shown in Chapter 3.3.1 formed the starting point for each of the Pipe and Breaks tables. Additional columns were added to incorporate the fields actually provided by the utilities. In many cases, utilities provided fields with different names but similar content. In these cases, the similar fields were all mapped to the same field in the common table. The aim of this practice was to generate predictive models which generalize across multiple utilities, including extrapolating to new utilities, rather than models which need to be trained separately for each utility. Data dictionaries summarizing the fields expected to be used in the modeling, their meaning, and how

many utilities included a field mapping to it are provided in Table 11 for the Pipes table and in Table 12 for the Breaks table.

Table 11: Data dictionary for Pipes table

Field	Meaning	Included
dataset	Unique identifier of each data provider	6
pipeid	Unique identifier of a pipe segment	6
from	PipeID of the upstream pipe segment	2
to	PipeID of the downstream pipe segment	2
dia	Nominal inside diameter of the pipe	6
mat	Pipe material	6
len	Length of the pipe segment	6
lining	Lining on the inside wall of the pipe	5
cor_protect	Corrosion protection applied to the pipe	2
joints	Joint type between pipe sticks	2
main_type	Water main type: distribution or transmission	4
deadend	Is this a dead-end, with no downstream segment	1
inst_date	When was the pipe installed	6
rem_date	When was the pipe removed	3
rehab_date	When was the pipe rehabilitated	3
active	Is the pipe currently in active use	3
loc1	Location: district, or service area	4
loc2	Location: postal code or pressure zone	2
owner	Owner: Utility, Customer, or Third Party	3
water_type	Potable water, raw water, or sewage	2
soil_type	Description of the soil type	2
note1	Free form text field for notes about the pipe	4

Table 12: Data dictionary for Breaks table

Field	Meaning	Included
dataset	Unique identifier of each data provider	7
brid	Unique identifier of a break event	7
pipeid	The pipe on which the break occurred	5
date	The date of the break	7
addr	Street address nearest to the break	5
loc1	Location: City, town, or district	5

loc2	Location: Postal code or county	3
event_class	Work order event type	3
break_type	Type of break (leak, burst, crack, etc.)	2
cause	Cause of break (split joint, corrosion, etc.)	3
soil_depth	Measured soil depth at break location	2
soil_type	Measured soil type at break location	3
xcoord	Longitude at break location	2
ycoord	Latitude at break location	2
pipe_mat	Pipe material at break location	3
pipe_dia	Pipe diameter at break location	3
note1	Free form text field for notes about the break	5

While additional fields were provided by utilities, these were present in only a single data set and did not represent factors expected to contribute to the analysis.

Certain fields required additional processing steps to populate for some utilities. Rather than providing one break per row, one utility provided one row per pipe, and columns for the date of the 1st, 2nd, 3rd, etc., through to the 9th break on that pipe. Some utilities omitted a unique identifier for each break record, requiring one to be generated. Some utilities stored the address in a single field whereas others had separate fields for number, street, city, and postal code. Some utilities spread the unstructured Notes associated with one break across multiple rows in the table. Many utilities included a range of maintenance records together with the actual break records, requiring the creation of a Boolean "isBreak" field populated by a utility-specific formula. Many utilities had multiple different date formats in the installation date field (e.g., "1921-03-07", "19210307", and "03/07/1921" all in the same utility's data), which needed to be made consistent. A similar issue was present in the pipe diameter field for many utilities, where units of measure were sometimes but not always included (e.g., "4", "4IN", and "4 inches"). Finally, some utilities included sentinel values that did not conform to the expected data type for the field, such as "July of 1982" for the installation date.

3.3.2.4 Data Integration Step 4: Merge Data from All Utilities into Pipe and Break Tables

This merging was accomplished by starting with an empty template table containing all of the target fields but no data, and inserting all of the rows from each of the individual utility data tables.

To facilitate easy analysis, date components were extracted from certain fields. For the Pipes table, the year was extracted from the installation, rehabilitation, and removal dates, and the month was also extracted from the installation date. For the Breaks table, the event date was broken down into separate year, month, day, and day of week values.

The net result of these operations were final Pipes and Breaks tables ready for data transformation and analysis.

3.3.3 Data Transformation

The data transformation process to prepare for analysis involved three main steps:

- Data Transformation Step 1: Creating segment-year tuples
- Data Transformation Step 2: Aggregating cohort-year tuples
- Data Transformation Step 3: Joining Pipe and Break data

These steps are described in detail to ensure repeatability of this research.

3.3.3.1 Data Transformation Step 1: Creating Segment-Year Tuples

The bulk of the planned analysis was conducted based on breaks occurring within a given time window on a given pipe segment or within a given cohort. Creating a list of valid segment-year tuples within which to populate data for analysis was a necessary step for this transformation. A given segment-year tuple was considered valid for this analysis if it occurred while the pipe was installed, and the utility was recording breaks for that pipe. This was calculated by ensuring all the following conditions were met:

1. The year was no earlier than the installation date of the segment
2. The year was no later than the removal date of the segment
3. The year was no earlier than the earliest recorded break for that data set
4. The year was no later than the last recorded break from that data set

Upon completion of this transformation, a total of 14,540,438 valid Segment-Year tuples were available for analysis.

3.3.3.2 Data Transformation Step 2: Aggregating Cohort-Year Tuples

It is common practice among utility asset managers to make rehabilitation and replacement decisions on the basis of cohorts. Cohorts can be defined in a wide variety of ways, such material, diameter, installation dates, geographic proximity, failure history, or the presence of corrosion protection measures such as linings. Grouping pipes into cohorts and making decisions based on these allows for more concise communication of plans and decisions. For example, "we will apply an internal cement mortar lining to all of our 1950s ductile iron mains and will replace all of the asbestos cement mains in our downtown area" is a clear and understandable plan, whereas a listing of hundreds of individual pipe identifiers cannot always be communicated concisely.

This concise communication can be critical to water utilities because of their regulated operations. Most water utilities are either public sector entities or operate under the purview of a public sector regulator. In either case, their capital expenditures are generally subject to an approval process, as these expenditures are a major driver of the water rates charged to the utility's customers. A clear and concise communication of which pipes will be replaced or rehabilitated and the rationale for doing so can help these plans to be explained both to the regulatory approvers and to the public.

For the purposes of this study, a simple and consistent method of defining cohorts was used. A cohort was defined as a combination of material, diameter, installation era, and geographic proximity. Grouping methods were applied across each of these four dimensions, as described below.

- **Material.** The same pipe material was often described with different words, short forms, or acronyms. For example, a cast iron pipe was described as CI, Cast, Cast Iron, Iron, Gietijzer (Dutch), Grijs Gietijzer (Dutch), GG (Dutch Acronym), among others. A rules-based script was used to standardize material names into the eight common water main material types: Cast Iron, Ductile Iron, Steel, Concrete, Asbestos Cement, PVC, Copper, and HDPE (including other Polyethylene types). Names which could not be clearly identified were grouped as Unknown.
- **Diameter**. Grouped into bins of 100mm (e.g., 0 – 99.99mm, 100 – 199.99mm, etc.).
- **Installation era**. Grouped by decade (e.g., 1950 – 1959, 1960 – 1969, etc.).
- **Geographic Proximity**. Grouped by utility.

3.3.3.3 Data Transformation Step 3: Joining Pipe and Break Data

At this stage in the processing, each Break record was associated with a single year and Segment. Some Segments and Cohorts had multiple breaks in the same year. As a result, the individual Break records had to be aggregated before joining the Break data with the Segment-Year or Cohort-Year tuples. The Break records are grouped using Dataset, Break_Year, and either PipeID or CohortID as the group keys, with the count of BreakID's representing the number of breaks on that segment or cohort respectively.

These aggregations were then joined with the list of valid Segment-Year and Cohort-Year tuples previously generated. A time-window operation was then used to generate additional break history features for a given Segment-Year or Cohort-Year tuple. Windows used include:

- Potential target variables <may include information from the future>:
    - A binary feature of whether any breaks occurred in the current year.
    - A binary feature of whether any breaks occurred in the five-year period beginning in the current year.
    - The number of breaks that occurred in the current year.
    - The number of breaks that occurred in the five-year period beginning in the current year.
- Potential predictive variables <may only include information from the past>:
    - The number of breaks that occurred in the prior year.
    - The number of breaks that occurred in the three-year period ending in the prior year.
    - The number of breaks that occurred in the five-year period ending in the prior year.
    - The number of breaks that occurred in all years up to and including the prior year.
    - A binary feature of whether any breaks occurred in any previous year.
    - A binary feature of whether any breaks in the five-year period ending in the prior year.

These form the final Segment-Year-Breaks and Cohort-Year-Breaks tables used to train and test the predictive models. The actual models are trained on subsets of these features, as described in the chapters covering the algorithm in question.

3.4 Exploratory Data Analysis

This section describes the exploratory data analysis and modeling performed prior to creation of the final model. The purpose of the exploratory work is to inform the design and selection of features, the choice of model types, and the selection of hyperparameters.

3.4.1 Data Preparation for Exploratory Data Analysis

While an objective of this model is to automate this process, a manual exercise of data preparation was undertaken as well. This was both to facilitate exploratory data analysis on clean data and to provide a version of the data against which the results of the automated processing could be compared.

3.4.1.1 Normalize the Units, Date Formats, and Data Types

This step required normalizing the date formats, units of measure, and data types across the various data sets into common standards. Numerical fields were all set to floating point decimal type with the exception of diameter, which was set to integer values. All numerical values were converted to metric units. All dates were converted from the text representation used by that utility into the PostgreSQL Date format.

The units used for numerical fields were as follows:

- Nominal Diameter: millimeters (rounded to the nearest integer when converted from inches)
- Length: meters
- Soil Depth: meters
- Roughness: microns
- Pressure Class: megapascals
- Tensile Strength: microns

- Measured Diameter (inside & outside): millimeters

3.4.1.2 Data Cleansing

Data cleansing scripts were prepared for both the Breaks and Pipes tables. Data cleansing activities prior to analysis were limited to those which each utility could reasonably be expected to be able to do on their own data, without access to data from other utilities. Given the large number of pipe segments (over 500,000) and breaks (over 180,000), a full manual inspection of all records for data quality issue identification and correction was not feasible.

As an example, rows were deleted where pipes had installation dates earlier than 1700 or after the date upon which the data sets were shared. Quite common were installation dates in the years 0 and 9999, both of which were likely sentinel values used to represent some specific but unknown information.

A more complex example was certain pipes having diameter records well below the normal minimum watermain size of 4 inches (100 mm). In some cases, this was clearly due to a units error during entry, such as a 1000 mm pipe being entered as 1 (length in meters) rather than 1000. These were corrected by identifying diameters for which water mains are not manufactured at that size (such as 1 mm) but which would exactly match a standard water main size if such a unit error is inverted.

3.4.2 Dataset Profiling

The analytical approach began with a data profiling exercise, consisting of exploratory data analysis and factor analysis, with a focus on the impact of individual factors. This aimed to validate and extend prior results cited in Chapter 2.2.2.2. It also served to verify the integrity and quality of the data set. The metrics tested include:

- Segment Length: Impact on annual failure rate (to confirm normalization per 100 km)
- Pipe Diameter: Impact on annual failure rate per 1,000 segments and per 100 km
- Pipe Material: Impact on annual failure rate per 1,000 segments and per 100 km
- Utility: Explore the variability of the data by utility
- Age: Impact on annual failure rate per 1,000 segments and per 100 km

Profiling by these factors ranged in complexity. The first two factors (age and diameter) were simple and straightforward to profile against.

Material was somewhat more complicated due to different languages, spellings, short forms, and acronyms being used to describe the same pipe material. This was accounted for by compiling a term mapping table to convert all commonly used terms to the common pipe materials described in Chapter 2.1.1.2, along with "unknown" where a mapping could not be found. It is also worth noting that the material records are tightly coupled with the age records. The earliest date of utilities' break tracking data ranged from 1959 in Toronto Water through to 2010 for Singapore PUB. Pipe ages, however, ranged back much further, in some cases to the late 1800s. Pipe material usage has changed over time; in the past 50 years cast iron pipe has fallen out of usage and PVC pipe has been introduced. As a result, the available break history on Cast Iron pipe is all for older pipe, whereas the break history for PVC pipe is all for newer pipe.

Profiling based on age was complicated by the risk of survivor bias (Wald, 1943). Anecdotally from discussions with the participating utilities, segments or even entire cohorts suffering from many breaks within the first few decades after their installation are often replaced or rehabilitated well before exceeding their design life. Conversely, segments or cohorts with few failures are often left in service for decades beyond their design life. This phenomenon could create either of two effects. First, it could dampen the rate of increase of failures as age increases, as the worst performing pipe groups are removed. This effect could conceivably even cause failure rates to decline after extensive replacement or rehabilitation of poorly performing pipes. As noted under material profiling, the older pipe also tends to be of different materials than the newer pipe.

Profiling based on past failures presents a data quality challenge. The break history often does not extend to the beginning of the pipe lifetimes. As a result, the actual number of prior breaks is not available for many of the pipes.

3.4.3 Exploratory Modeling

Following the data profiling exercise, conventional machine learning models were applied to the data to provide baseline performance metrics and to select the model type for the generalized model.

3.4.3.1 Target Variables and Performance Metrics

The target variables used for the baseline models were as follows:

- For Segment-Year tuples: Binary classification on whether any breaks occurred that year.
- Alternate: Binary classification on whether any breaks will occur in the next five years.

The Segment-Year tuples dataset was highly imbalanced, with non-break years accounting for 99.1% of the data. With such a highly imbalanced data set, using accuracy as the performance metric would be ineffective, as a simple "always classify as not a break" model would be 99.1% accurate. The alternate five-year-horizon Segment-Year dataset was also highly biased, with 96.0% of the data reflecting a non-break period. To account for this class imbalance, all training samples were given weights inversely proportional to their "break / no-break" class frequency. This ensured each class received equal weight during training.

The most common practical use of the results of pipe failure prediction models such as these is to select a small subset of the pipe network as candidates for a replacement or relining program aimed at reducing the overall system break rate. Performance metrics have been selected with this in mind. An effective classifier would be one which captures a high percentage of the breaking segments among a relatively small number of predicted failing segments. Two common metrics that support this are the area under the curve (AUC) for the receiver operating characteristic (ROC), and the Cumulative Lift at a given low percentage.

The alternative target variable of whether a break occurred in the upcoming five years was also tested in consideration of this likely use case. Pipe replacement and rehabilitation programs generally take several years to plan and execute, so this may prove to be a more practical time horizon. The same features, models, and hyperparameters were used for this test as well.

Hyperparameter tuning was performed based on the AUC performance metric. This ensures that the selected model will perform well across a range of different percentages of the pipe population selected for replacement or rehabilitation. Performance of the selected models will be reported using Cumulative Lift at 10% of the population. This will be expressed as the percentage of all actual breaks captured if 10% of the system were to be selected for replacement or rehabilitation. This provides an easy and intuitive comparison metric for readers who do not have a background in machine learning.

3.4.3.2 Feature Selection and Preparation

For the Segment-Year tuples baseline model, the features selected, and the preparation via standardization, which involves subtracting the mean and dividing by the standard deviation (Bonakdari et al., 2023), or dummy encoding for each, were as follows:

- **Diameter**: Mean / Standard Deviation Rescaling
- **Material**: Dummy encoding, max of 10 categories
- **Length**: Mean / Standard Deviation Rescaling
- **Lining**: Dummy encoding, max of five categories
- **Corrosion** Protection: Dummy encoding, max of five categories
- **Joints**: Dummy encoding, max of five categories
- **Main Type**: Dummy encoding, max of five categories
- **Installed Year**: Mean / Standard Deviation Rescaling
- **Age**: Mean / Standard Deviation Rescaling
- **Rehabilitation Date**: Binary variable to flag presence

3.4.3.3 Selection of Train / Test Split

Standard best practice for training machine learning models is to separate the available data into three groups: training data, validation data (for hyperparameter tuning), and test data (for final evaluation). This helps avoid overfitting by ensuring that hyperparameters are not tuned on the test data. Unless otherwise noted, the test set was selected randomly from the training set, the top hyperparameters were selected using the validation set only, and performance metrics shown are on the test set by the model with the top hyperparameters.

Careful consideration must be given to the choice of the training and test data. A simple random assignment of Segment-Year tuples would introduce a substantial risk of information leakage from the future into the past. In the most extreme example, the training data could include future records from the same segment in the test data. In less extreme examples, the training data could include records from the same year for other pipe segments in the same cohort. Furthermore, random assignment does not establish whether the model extrapolates well to new utilities. Each utility has

certain unique descriptive values in the non-numerical fields, such as material and joints. Consequently, models have an opportunity to learn predictions specific to each utility. This would be particularly true for decision-tree based models such as random forests or gradient boosted trees, which can explicitly segregate predictions based on any parameter. A truly robust model should be able to predict failure rates for data from a new utility.

Each of these challenges was addressed through careful selection of training and test data splits. The information leakage challenge was overcome by excluding the most recent two years from the training data and using only the most recent two years in the test data. The generalization to new utilities was tested by using K-Fold Cross-Validation (Lu, 2010), applying a variant thereof wherein one fold is created for each utility. For each utility, the models were trained on all other utilities' data and then tested on that utility's data. Combining time splitting and Utility Cross-Validation approaches offers a stringent test of the generalization of a model. It is trained on past data from other utilities and then tested on future data from the target utility.

3.4.3.4 Models and Hyperparameters Tested

For the Segment-Year classification model, a selection of four common model types was tested, with a range of hyperparameters used for each model. All classifiers have been trained in Python, with the Random Forest, Logistic Regression, and Decision Tree classifiers from the Scikit Learn package, and a gradient boosted tree classifier from the XGBoost package. The models and hyperparameter values tested were:

- Random Forest (100 trees)
    - Maximum depth of trees: 4, 6, 8, 10, and 12
    - Minimum samples per leaf: 1, 2
- Logistic Regression
    - L1 Regularization with C of 0.01, 0.1, 1, 10, and 100
- Gradient Boosted Trees (XGBoost implementation; 100 trees; 10 round early stopping)
    - Maximum tree depth: 2, 3, 4, and 5
- Decision Tree (Gini split criterion, maximum tree depth of 5)

Hyperparameter tuning was conducted via 5-fold cross validation on a sample drawn from the training set, with the test set being reserved for final evaluation of the selected hyperparameters. Where multiple different hyperparameters were being tested, a full grid search was conducted across all combinations.

3.5 Generalized Machine Learning Model for Pipe Failure Prediction

The general problem statement provided in Chapter 2.2.6 is to provide an estimator $f(t, \boldsymbol{x})$ and a calibration factor c that allow estimation of the Probability of Failure (PoF) and Expected Number of Failures (ENoF) for a given pipe in a given time period t. Specifying a time period *t*, as recommended by Robles-Velasco et al. (2023), will allow the model to provide an estimate of the failure rate. This ensures that the model outputs can be used in a wide range of engineering applications (Scheidegger et al., 2015) and is often sufficient for making management decisions (Barton et al., 2022b). The additional forecasting of the expected number of failures, as introduced by Aslani et al. (2021), will allow for simple aggregation to cohorts, areas, or an entire utility, removing the need for a separate estimate of number of failures.

The objective of our model is to provide estimators for the Expected Number of Failures and the Probability of Failure in a manner that is fit to data in a training set. This is expressed by supplementing the equations for ENoF and POF in Chapter 2.3.1.1 with a parameter vector $\boldsymbol{\theta}$.

$$ENoF(t \mid \boldsymbol{x}, \boldsymbol{\theta}) = c \cdot PoF(t \mid \boldsymbol{x}, \boldsymbol{\theta})$$

$$PoF(t \mid \boldsymbol{x}, \boldsymbol{\theta}) = f(t, \boldsymbol{x}, \boldsymbol{\theta}) \quad \textbf{(30)}$$

Where:

- $\boldsymbol{\theta}$ is a vector of parameters
- *t* is the first year of a five-year period
- $\boldsymbol{x}$ is a feature vector describing a single pipe
- $ENoF(t \mid \boldsymbol{x}, \boldsymbol{\theta})$ is Expected Number of Failures in the five-year period beginning at *t*, for the pipe described by feature vector $\boldsymbol{x}$, using the parameters in $\boldsymbol{\theta}$

- $PoF(t \mid \boldsymbol{x}, \boldsymbol{\theta})$ is probability that a pipe described by feature vector **x** will experience one or more failures during the five-year period beginning at t, using the parameters in $\boldsymbol{\theta}$
- c is a calibration constant to convert between probability of failure and number of failures
- $f(t, \boldsymbol{x}, \boldsymbol{\theta})$ is an estimator function

Chapter 2.3.1.1 provides a general problem formulation for failure prediction problems, with the final equations also shown below. These equations describe the method that will be used to train a machine learning classifier as the estimator function $f(t, \boldsymbol{x}, \boldsymbol{\theta})$.

$$g(\mathbf{x}) = \begin{cases} 1 & \text{if} f(\mathbf{x}, \theta) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = P(y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta})$$

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{n=1}^{N} l(g(\boldsymbol{x}_n, \boldsymbol{\theta}), y_n) \quad \textbf{(31)}$$

Where:

- $\boldsymbol{\theta}$ is a vector of parameters
- N is the total number of samples in the training set
- $\boldsymbol{x}$ is a vector of features, with $\boldsymbol{x}_n$ describing the n^{th} sample in the training set
- y_n is the correct label for the n^{th} sample in the training set
- $g(\boldsymbol{x}, \boldsymbol{\theta})$ is the label assignment function as shown in Equation (24)
- $l(\hat{y}, y)$ is the loss function for predicted label $\hat{y}$ and correct label y, used to fit $\boldsymbol{\theta}$ to the training data

Note that while the class assignment function g() is necessary for fitting $\boldsymbol{\theta}$, it is not actually required in the final application. Our model outputs will be the probability estimates, enabling various sorting methods and decision thresholds to be used for different applications.

3.5.1 Model Structure

A three-layer structure illustrated in Figure 10 is used in the model to accommodate the requirements to generalize across utilities and applications, as set out in Chapter 2.2.6. The first layer is a pre-processing layer to allow the model to use data from a variety of sources as provided by the utility, irrespective of units of measure, language, jargon, and record-keeping decisions regarding what constitutes a single pipe. If effective, this will remove the need frequently stated in the literature for extensive data cleansing. The second layer is the core machine learning model, which provides as outputs both the class label (failed / not failed) and the probability assignment for each class. The final layer calibrates the relative probability estimates provided by the machine learning model and calculates the expected number of failures.

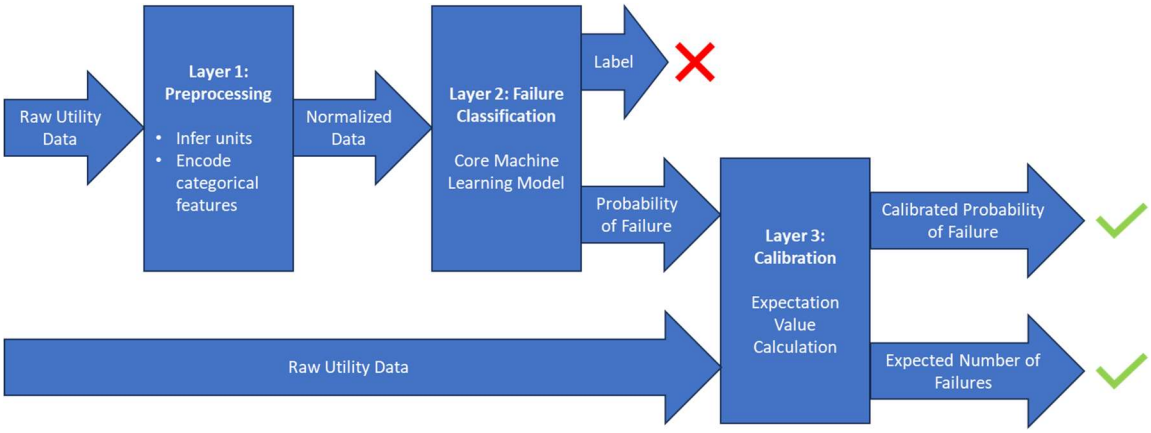


Figure 10: Structure of generalized machine learning model for pipe failure estimation.

Drawing concepts from the field of transfer learning, machine learning failure classification model is frozen upon training. No information from the utility being tested is permitted to be used in training this model. Layer 1 (Preprocessing) and Layer 3 (Calibration) are adapted on a per-utility basis.

3.5.2 Model Input: Raw Utility Data

A member of the population for this model (i.e., a single feature vector x) is a Segment-Year tuple. This may also be referred to informally as a Pipe-Year tuple.

As noted in Chapter 2.1.1.1.2, record-keeping decisions as to how pipe sticks were grouped into segments varied from utility to utility. This is evidenced by the varying average length of a pipe segment across utilities, as shown in Chapter 4.1.1. This variability is accepted as part of this study, with the definition of a segment being "whatever the utility tracks as a single record in its pipe database." Calibration for utility record-keeping practices is performed in the feature pre-processing described in Chapter 3.5.3.

Segment-Year tuples were created by joining the Segment and Break tables, as described in detail in Chapter 3.3.3.1. The specific training and test sets used are described in Chapter 3.6.

The features included from each Segment-Year tuple are those that were present in at least half of the participating utilities:

- **Diameter**: Nominal diameter of the pipe, in whatever units the utility generally uses
- **Material**: Text field describing the pipe's material of manufacture
- **Length**: Recorded length of the pipe segment, in whatever units the utility generally uses
- **Lining**: Text description of the internal lining applied to the pipe
- **Corrosion Protection**: Text description of external corrosion protection applied to the pipe
- **Joints**: Text description of the type of joints between pipe sticks
- **Main Type**: Text description of the type of main (distribution, transmission, etc.)
- **Installed Year**: Year of installation of the pipe
- **Age**: Age of the pipe during the target year (note that this is a calculated feature, where Age = Year of Interest – Installed Year)
- **Previous Rehabilitation**: Binary field indicating whether the pipe record showed a Rehabilitation Date prior to the target year
- **Breaks in [1, 3, 5] Prior Years:** Numerical fields indicating the number of breaks in the [1, 3, and 5]-year period prior to the year of interest.

These features have been selected to conform to the requirements set out in Chapter 2.2.6 for the feature vector x. A short time horizon was selected for the prior breaks features, to reduce the

chances of unavailable data for new utilities, and also to mitigate the risk of left censorship of training data.

The features selected will also permit running the model for a target time t that is further out into the future. Most of the features selected are permanent aspects of the pipe. Only the Age, Previous Rehabilitation, and Breaks in [1, 3, 5] Prior Years are time dependent. These can all be updated simply if a future point in time t is to be simulated. Age can simply be calculated using the Installed Year. The value of Previous Rehabilitation can simply be drawn from the most recent value for that pipe. The values of the Breaks in [1, 3, 5] Prior Years can be left as per the most recent values.

The target variable for the machine learning model is whether any breaks occur in a Segment-Year tuple in the upcoming five-year period. As discussed in Chapter 2.2.6, the five-year time horizon was chosen to provide the most practical application value. This time horizon matches the typical time horizon used in utility Capital Improvement Planning processes. It also mitigates the challenge of an unbalanced data set.

The definition of a break varied from utility to utility among those contributing to the study. No specific definition of a break was required beyond "whatever each utility considers a break." Calibration for this aspect of the utility record-keeping practices is performed by Layer 3 in the model, as described in Chapter 5.3.

Cohort-Year tuples were used to confirm that aggregation of the model results facilitates forecasting the number of failures in a cohort.

3.5.3 Layer 1: Feature Preprocessing

As described in Chapter 4.1.2, a major challenge with integrating data from various contributors is variability in the data contained in what is nominally the same field. While not extensively reported on in the literature, this challenge is likely present in many other domains where data needs to be integrated from a variety of contributors. This preprocessing layer aims to address this challenge in a manner that may be applicable to data integration across disparate sources for any machine learning classification problem.

A general approach to integrating data from various sources is presented below. It assumes only the data type integrity for numerical, date, and text features. Processing steps are as follows:

3.5.3.1 Infer Units for Objective Numerical Features

Objective Numerical Features represent an objective, real-world quantity, such as pipe diameter. The units used will vary across data contributors and need to be inferred. For each type of units (such as length), there are a finite number of different possible units used by each utility (e.g., mm, cm, inches, feet, meters, km, or miles for a length feature). The most likely units for each dataset are inferred using aggregate statistics of that feature for that dataset. Units were then normalized to a consistent standard.

Aggregate statistics used were mean, median, and standard deviation. These were selected as they are measures of the scale of the variable, which would scale linearly with the values. This allows possible unit conversions to be tested directly on the aggregate values.

For each objective numerical feature, the plausible units of measure were listed. Conversion factors to a desired standard unit were prepared in a list. All possible combinations from these unit conversions among the utilities contributing to the sample set were listed. For each combination, the resulting converted mean, median, and standard deviation were listed for each data contributor. The scaled variance (i.e., the variance divided by the mean value) across data contributors was calculated for each aggregation (mean, median, and standard deviation), providing measures of the similarity of the converted values. The scaling of the variance was selected to avoid introducing a bias in favor of smaller values (e.g., the unscaled variance with all units in km would be smaller than the unscaled variance with all units in m for the same data). The sum of these three measures was taken. The combination with the lowest sum was selected as the correct unit conversion combination.

For the selected combination, the mean of each aggregation across utilities was stored as a model parameter. At test time, when a new data contributor was added, each metric was calculated for the new utility and compared to these averages. The unit conversion which minimizes the sum of these differences is selected for application to that utility at test time.

For the specific case of pipe condition assessment, the plausible units of measure were:

- Pipe Diameter: mm, inches
- Pipe Length: feet, meters

A further constraint was applied that both Diameter and Length must be either metric (mm, meters, km) or imperial (inches, feet, miles) for each utility. The sum of the scaled variances for the final

units selection was taken jointly across the two measures. It is noted that with a larger number of plausible unit pairings, the number of combinations to test would become quite large. For example, if there were four plausible sets of units (two metric, two imperial) for each variable, this would leave eight possible unit pairings for each utility, with six participating utilities, for 8^6 (262,144) possible combinations to be tested. The requirement previously mentioned that aggregations scale linearly with unit conversions was put in place for computational tractability, given the large number of combinations to test.

To ensure generalizability forward in time, only data from the training set was permitted for use in inferring units for numerical features. Since the training set data was strictly prior to the test set data for any given utility, this ensures that there is no information leakage from the future into the past in this step. The unit conversion factors, as inferred using the training data, were then applied to both the training and test data for each data contributor.

3.5.3.2 Normalize for Subjective Numerical Features

Some numerical features involve an element of human judgement or differences in administrative record-keeping processes. An example is pipe break frequencies. There is some ambiguity as to what constitutes a pipe break. For example, some utilities own the service lines connecting each customer to the network and would consider a leak on these a main break, whereas for other utilities these may be the responsibility of the customer and not recorded as a break in their database. Features of this type will have the per-contributor average value added to the original feature vector. This allows the machine learning model in Layer 2 to discover appropriate normalization relationships between the tendencies of data contributors, the individual variable values, and the target variable.

As an additional normalization step, one further feature is introduced to the feature vector. Based on the strong inverse relationship between pipe diameter and failures per unit length described Chapter 4.2.1, the length to diameter ratio was added as a feature. This feature also provided an additional hedge against the possibility that a particular utility uses units on a different scale from others (e.g., using m for pipe diameter and km for length).

3.5.3.3 Encode Categorical Features

The standard method of handling categorical features is to create dummy variables. This method fails to extend to new datasets which contain new labels not previously seen. This is common among utilities, where different languages and jargons will be in use. For example, the terms “Grey Iron”, “Centrifugally Cast”, and “GG” all refer to the same physical type of pipe in our data set.

The potential for encoding or embedding these categorical features has been noted in past literature reviews (Delnaz et al., 2023). A simple approach was proposed (Robles-Velasco et al., 2020) whereby each categorical variable is replaced by ordinal encoding when the labels are sorted by their respective average failure rate per km per year.

We introduce a more complex encoding method, called cross-encoding, with the aim of generalizing across data contributors. Instead of dummy encoding, each label pairing was replaced with a vector representing statistics about numerical features from the training data where the entries had that value. For example, the pipe_mat field may be replaced with [pipe_mat_dia_avg, pipe_mat_install_year_avg, pipe_mat_len_avg, pipe_mat_breaks_per_1000_seg_year_avg]. Since these vectors are constructed after the units inference step, the same physical pipe materials should be represented with similar vectors. Continuing our example, the entries for different types of cast iron would likely be close to [150, 1940, 80, 50], whereas entries for concrete pipe would be closer to [500, 1970, 300, 5]. To allow for different practices among data contributors, aggregate values were calculated for each feature-contributor tuple (in this study, each feature-utility tuple).

The output of the preprocessing layer is the feature vector which will be passed to the machine learning model. Note that there is no further preprocessing of categorical or numerical variables. No rescaling is needed for numerical variables, as they have already been normalized to the same units, and the decision-tree based models used in Layer 2 are not sensitive to feature scale. Since all categorical variables have already been encoded, there is no need to create dummy variables.

3.5.4 Layer 2: Failure Classification

This layer employs a machine learning classifier, labelling Segment-Year tuples based on whether they will experience a break in the upcoming five-year period. Any machine learning classifier can be used, provided that it conforms to the following requirements:

- Able to capture non-linear relationships with and among features.

- Able to provide a probability estimate for the “Break” class.
- Is relatively robust to missing (null) values in some features.

The particular model selected for use in this study was gradient boosted trees. This decision was informed by both the literature review, where several studies have indicated strong performance by gradient boosted trees on pipe failure assessment problems (Delnaz et al., 2023), as well as by the observations in the Exploratory Modeling exercise in Chapter 4.3.

Hyperparameter selection was performed based primarily on the results obtained during the exploratory modeling described in Chapter 4.3.1.6. The exception was the maximum number of trees and early stopping. In the exploratory modeling, a maximum of 100 trees was used, with early stopping occurring if no improvement in the loss function was achieved in 10 rounds of adding a tree. This setting caused repeatability problems with the final model implementation described below. As such, a fixed value of 40 trees was used, with early stopping disabled. The value of 40 trees was selected as a typical value at which early stopping occurred during the exploratory modeling. For replicability, the full setting string used to train the model is provided below.

```
Xgb_params = {
        'max_depth': 4,
        'learning_rate': 0.2,
        'gamma': 0.0,
        'min_child_weight': 1.0,
        'max_delta_step': 0.0,
        'subsample': 1.0,
        'colsample_bytree': 1.0,
        'colsample_bylevel': 1.0,
        'reg_alpha': 0.0,
        'reg_lambda': 1.0,
        'n_estimators': 40,
        'silent': 0,
        'nthread': 4,
        'scale_pos_weight': 1,
        'base_score': 0.5,
        'random_state': 1337,
        'missing': None
    }
```

To account for the imbalance between the “Break” and “No Break” classes, samples were weighted with inverse proportion to the number of samples in the training set with that class.

Three versions of the model were trained and tested in each instance, to help with assessing of the effectiveness of the full model in generalizing. The models were as follows:

- **Isolated**: Six estimators were trained, one for each utility contributor using only the data provided by that contributor. At test time, each utility was evaluated using the estimator trained with its own data. This simulated the scenario wherein a single utility uses their own data in isolation to predict their own future breaks.
- **Inclusive**: One estimator was trained using all the training data from all six utility contributors. At test time, each utility was evaluated using the same estimator. This simulates the scenario wherein multiple utilities pool their data to train a common model.
- **Leave One Group Out (LOGO)**: Six estimators were trained, one for each utility contributor using only the data provided by other contributors (i.e., excluding that utility's data in the training set). At test time, each utility was evaluated using the estimator trained without using its data. This simulates the scenario wherein a pretrained model is used by a new utility that did not contribute data to the model.

Each version of the training simulates a different practical situation. The Isolated situation represents each utility training their own model on their own data. The Integrated situation represents multiple utilities choosing to pool their data on an ongoing basis, for the sake of training a model which all could use. The LOGO situation represents a group of utilities choosing to pool their data on a one-time basis, with the resultant model being used by other utilities.

The LOGO model is proposed as the generalized model by this study. The experience of the authors in gathering data for this study suggests that expecting utilities to pool their data on a regular basis is impractical. Furthermore, few utilities have both sufficient failure records and the capacity to build and train their own models. The scenario where a model is trained once and then used broadly is by far the most practical.

As previously noted in Chapter 3.5.1, the class label assignments created by the machine learning model are dropped at this stage. Only the probability of the "Break" class, $PoF(t \mid \boldsymbol{x}, \boldsymbol{\theta})$ is used by the final layer.

3.5.5 Layer 3: Calibration

The first step in calibration is to calibrate the probability estimates f(x, θ) output by the XGBoost binary classifier. These probability estimates are tuned to maximize the application of correct class labels, rather than to provide correct probability estimates, and as a result may not be well calibrated to the true probabilities. This is a common issue when classes are imbalanced (as was the case for these models), and can be resolved by calibrating the probability estimates (Niculescu-Mizil & Caruana, 2005). The degree to which the predicted probabilities match the actual observed probabilities is summarized by the Calibration Loss metric (Kull & Flach, 2015).

Isotonic Regression was used to calibrate the XGBoost output probabilities to the actual observed probabilities. This is a nonparametric regression calibration, which is done by grouping the samples into bins of similar estimated probability and calculating a calibration factor per each bin which aligns the average estimated probability in the bin with the average observed probability in the bin (Zadrozny & Elkan, 2002).

The final step in the calibration layer is converting the Probability of Failure (PoF) estimate to an Expected Number of Failures (ENoF) estimate. This conversion accounts for the possibility that multiple failures may occur on the same pipe within the same time window (five years, in the case of this model). A simple multiplicative linear calibration is applied on a per-utility basis. The calibration was calculated using only the two most recent years of data in the training set for each utility. The rationale for this was twofold. First, it simulates a utility with only a limited break history using the model. Second, it accounts for the possibility that a utility with a longer break history may have changed its record keeping practices from time to time, with the most recent data more likely to reflect their current practices.

3.6 Evaluating the Model to Confirm Generalizability

This section elaborates on the concepts of generalizing forward in time, and generalizing across different utilities, regions, and languages.

3.6.1 Defining the Train / Test Split to Generalize Forward in Time

The train and test data were split in a manner to ensure that the model extrapolates forward in time. For each utility data contributor, the most recent seven years from their data was reserved as the test set. This allows for three target years, each with a full five years of recording data to ensure no right-

truncation of the target variable. All other data is designated for the training data set. This ensures a long history is included in the training data, reducing the chances that the model will overfit on a short-term phenomenon.

As with the exploratory data analysis, only a sample of the data could be used due to memory limitations. A random sampling of 1,000,000 records (800,000 for training, and 200,000 for test) was employed. The samples were drawn equally from the contributing utilities.

3.6.2 Leave One Utility Out Cross-Validation to Generalize to New Utilities

Model evaluation was performed using K-Fold Cross-Validation (Lu, 2010), applying a variant thereof wherein one fold is created for each utility. For each utility, the models were trained on all other utilities' data, and then tested on that utility's data. This approach, known as "Leave One Group Out" cross validation, was used by Daulat et al. (2024) as a method of testing the ability of a model to extrapolate to new utilities. This approach simulates the situation where all but one of the utilities contributed data to train the model, which was then applied to the new utility. This reflects the scenario wherein a pre-trained model being used as-is (with no retraining) by new utilities.

Of particular interest is how this model will perform when tested on a utility that employs a language or units of measurement which were not previously seen in the training set. This will provide the strongest confirmation of the generalization to new utilities.

When combined with the train/test time splitting noted in Chapter 3.6.1, this provides a reasonable estimate of the performance of the model were it to be used by a new utility, at a future point in time, the problem statement described in Chapter 2.2.6.

3.6.3 Selection of Performance Metrics to Generalize Across Applications

This section discusses the evaluation structure used to compare performance of different models, including a discussion of the various performance metrics used in the literature.

As noted in Chapter 3.5, while the class assignment function g() is necessary for fitting θ, it is not required in the final application. Consequently, the choice of threshold (here 0.5) is not relevant to this study. Any performance metrics which depend on the threshold (such as accuracy, sensitivity, specificity, etc.) were avoided. However, performance metrics which aggregate across possible values of the threshold remain valuable.

Two primary performance metrics have been selected for this study. The first is the Area Under the Curve of the Receiver Operating Characteristic. The second is Lift at 10%. Each has been selected for its alignment to the specific decisions to be addressed by this study, as described in Chapter 2.1.4.

The Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC) is a widely used metric in machine learning for evaluating the performance of classification models. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various decision threshold settings. The AUC represents the degree to which the model can distinguish between classes. An AUC of 1 indicates perfect classification, while an AUC of 0.5 suggests no discriminative power, equivalent to random guessing (Fawcett, 2006). The mathematical formulation of AUC can be represented as the integral of the ROC curve:

$$\text{AUC} = \int_0^1 ROC(\text{fpr})\ d\text{fpr} \tag{32}$$

Where:

- fpr = False Positive Rate
- $ROC(\text{fpr})$ is the value of the Receiver Operating Characteristic at a given false positive rate

A higher AUC indicates a better model performance with a greater ability to differentiate between the positive and negative classes, with an AUC of 0.5 being the score of a "random guess" classifier and an AUC of 1.0 being the score of an oracle (a classifier which is always correct).

A plot of the Receiver Operating Characteristic provides a good representation of the effectiveness of a binary classifier at various selections of the classification threshold. In the case of pipe failure classification, the x-axis corresponds to the proportion of replaced pipes which would not in fact have incurred a break during the five-year period, whereas the y-axis corresponds to the proportion of the currently anticipated breaks which could be avoided in the upcoming five-year period by replacing those pipes. Each potential value of the decision threshold represents a different level of investment in a pipe replacement program. A lower threshold corresponds to targeting more pipes for replacement, and a higher threshold corresponds to targeting fewer pipes. The ROC curve hence provides a visual representation of the effectiveness of a range of different levels of investment in

pipe replacement. The Area Under the Curve provides an aggregate metric of this performance. The downside of the AUC performance metric is that it is not intuitive and easy to explain to pipeline management professionals.

Lift, also known as cumulative lift, is a machine learning metric used to assess the effectiveness of a predictive model, particularly in binary classification problems. It measures the model's ability to identify positive instances (e.g., a rare event) compared to random chance. Lift at a low percentage is appropriate for use in imbalanced data sets because it focuses on the early detection of positive instances, which is often more critical in such scenarios (Lu, 2010). The formula for calculating lift at a specific percentage x, is as follows:

$$\text{Lift}(x\%) = \frac{\text{Percent of True Positives Cases in Top } x\%}{\text{Percentage of Positive Cases in data set}} \quad \textbf{(33)}$$

Where:

- Percentage of True Positives in Top x% = the proportion of actual positive instances correctly identified by the model among the top 10% of predictions
- Percentage of Positive Cases in data set = the overall proportion of positive instances in the dataset

Lift(x%) has a simple and intuitive conceptual explanation. Among the x% of samples with the highest estimated probabilities, how many times better did the classifier do at correcting picking positive samples than random guessing would have? In the context of pipe replacement, this can be explained as "if you select the x% highest risk pipes, this includes <Lift(x%) * x%> of the breaks."

Lift at 10% was chosen as an intuitive performance metric, which is well aligned to the primary target application of selecting pipes for replacement. This metric provides the proportion of pipe failures in the upcoming five-year period which would be averted by replacing the 10% of pipes deemed by the model to be at highest risk of failure. For example, if the Lift at 10% is 4.5, this means that replacing this 10% of the pipes in a network would avoid 45% (4.5 * 10%) of the pipe breaks for the ensuing five-year period. Along with being intuitive and easy to explain, Lift at 10% directs the focus of the performance evaluation to the most at-risk pipes. The practical value of measuring performance when selecting a small number of pipes for replacement was noted by

Robles-Velasco et al. (2020), who manually tuned decision thresholds to measure the percentage of breaks that could be avoided by selecting a small percentage of the pipes. The Lift at 10% metric generalizes their approach.

Further metrics were selected for evaluating the performance of the full model including the calibration layer.

Log Loss and Calibration Loss were used to measure the effectiveness of the Isotonic Regression calibration step. Log Loss is impacted by both the relative probability estimates between samples and the alignment of probability estimates to actual observed probabilities, whereas Calibration Loss measures only the alignment of probability estimates to actual observed probabilities.

For the final calibration from Probability of Failure (PoF) to Expected Number of Failures (ENoF), cohorts were formed as follows. Each utility – material – diameter grouping was considered a separate cohort. Diameter groupings were based on the following bins (all sizes in mm): [0 – 99; 100 – 199; 200 – 299; 300 – 399; 400 – 599; 600 – 899; 900+]. Any cohorts too small for meaningful analysis (i.e., experiencing fewer than one failure per year on average) were dropped. Performance was measured by comparing the ENoF summation to the actual total number of breaks in cohort over the ensuing five-year period.

As a reference point, the performance of a simple "Predict the same number of breaks as occurred in the previous five-year period" model (referred to as the Prior Period model) was also measured. The following metrics are presented:

- Mean Absolute Error (MAE)
- Bias
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMS error, or RMSE)
- Linear Regression Best Fit (coefficients closer to 1, and intercepts closer to 0, are better)
- R^2 Value for linear regression best fit line

3.7 Chapter Summary

This chapter provided a detailed methodology for the research project. It described the four principle steps of the project: (1) collecting data from utilities, (2) integrating it into a standard data model for analysis, (3) exploratory analysis and modeling, (4) creating a generalized model for pipe failure analysis, and (5) evaluating the performance of this model. The three-layer structure of the generalized model was described in detail, with a focus on the novel methods for inferring units and encoding categorical features to remove the need for manual data cleansing.

The following chapters describe the results obtained by following this methodology. Chapter 4 describes the results of the exploratory analysis, and Chapter 5 describes the results of the generalized machine learning model for pipe failure prediction.

Chapter 4
Results: Exploratory Data Analysis

Exploratory data analysis has been performed on the Segment-Year tuples generated from the six utilities providing fully usable data: City of Toronto, Peel Region, Hamilton, American Water, Waternet, and Singapore Public Utilities Board (PUB). The factors investigated and preliminary results are described below.

4.1 Dataset Profiling

This section provides summary statistics of the dataset.

4.1.1 Basic Descriptive Statistics

The overall dataset consists of data from six utilities, with 153,868 breaks on 581,974 pipe segments spanning 31,535 km of pipe. The break recording periods by each utility range between 13 and 56 years. These monitoring periods result in a total of 10,046,561 segment-year tuples, accounting for 692,732 km-years of pipeline failure records.

Detailed statistics from the key data tables are available in Appendix A.

4.1.2 Exploration of Distribution of Data

Figure 11 shows the distribution of pipe length currently in service (i.e., not abandoned) and total recorded breaks among different materials, with the top 20 shown for each. Cast Iron is both the most used pipe material, and the largest contributor to the total break count.

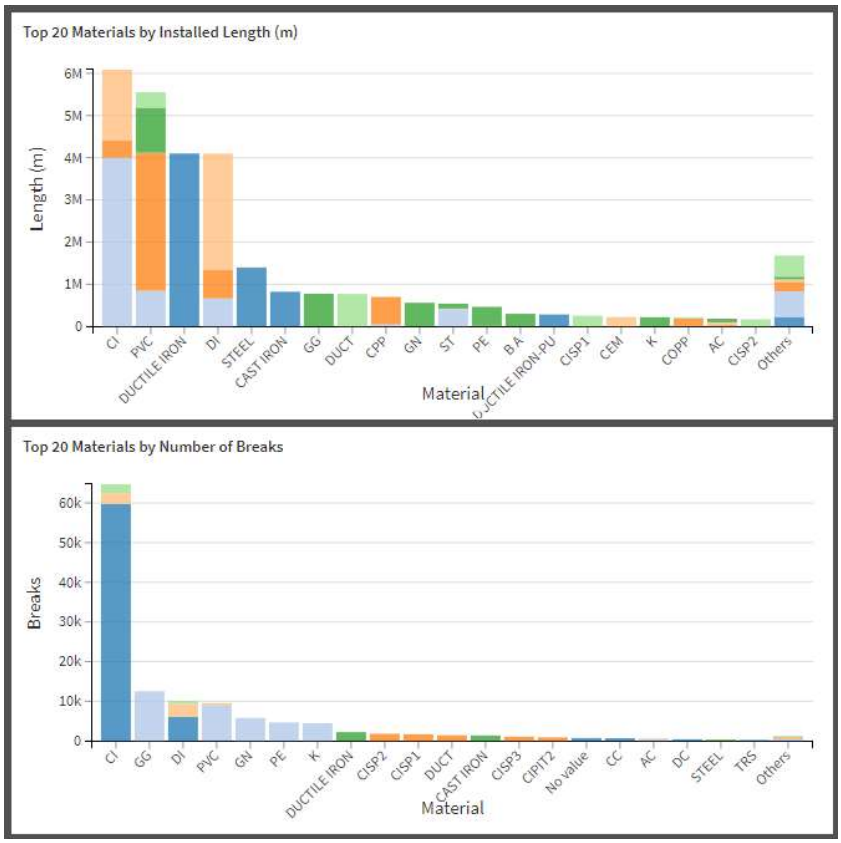


Figure 11: Top materials by in-service length (top) and number of breaks (bottom).

Many of the material names used by the utilities are either duplicative (such as CI and CAST IRON) or not standard English pipe material acronyms (such as GG). The bulk of these names have been normalized to the eight common water main materials for profiling purposes. This normalization was accomplished through careful manual review of the data, considering the diameter, installation year, and the local language and terminology. For example, in the Netherlands, cast iron is commonly referred to as "grey cast iron" which translates to "gietijzer grijs" or GG for short. Likewise, ductile iron is referred to as "Cast Iron Ductile" which translates to "gietijzer nodulair" or GN for short. These translations were confirmed via the diameters and installation dates of the pipes in question. Figure 12 shows the total length of installed pipe and number of breaks after this material name normalization has been applied. After normalization, Ductile Iron is now the most common material; however, the majority of breaks still occur on Cast Iron.

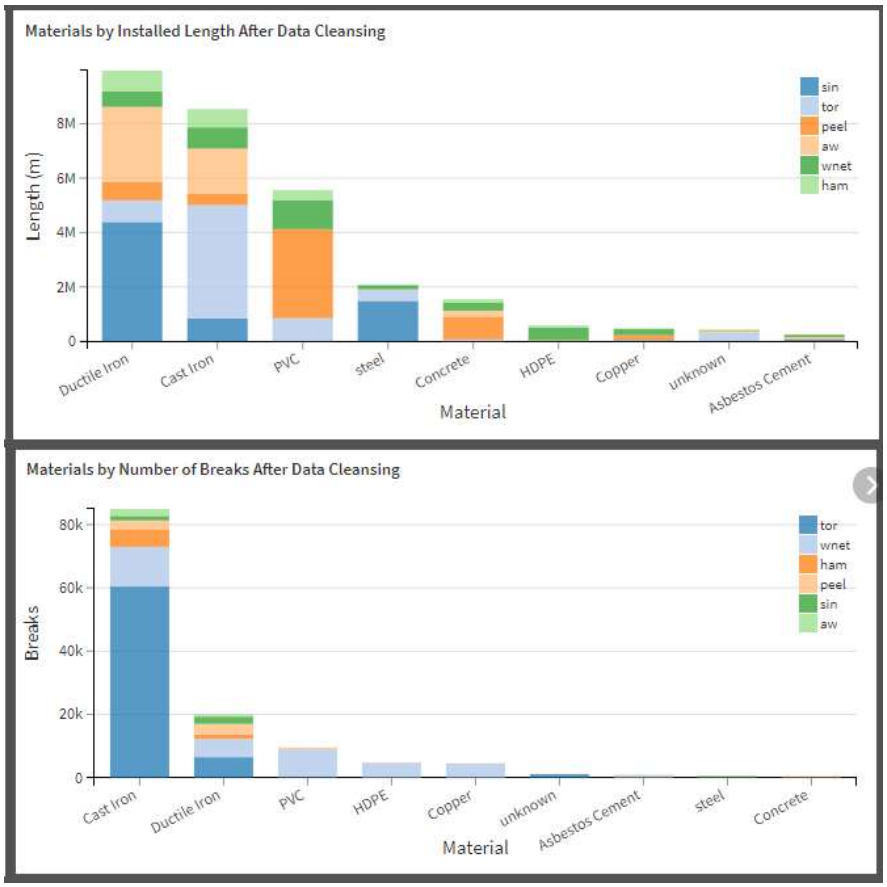


Figure 12: In-service length (top) and number of breaks (bottom) by normalized material type.

One reason for the high number of breaks recorded in cast iron pipe is apparent from the distribution of pipe installations over time, as presented in Figure 13. Installation years are binned into groups of five years. This was to account for a suspected data quality issue, wherein the majority of older (pre-1960) installation data were shown in years that were exact multiples of 5 (e.g., 1955, 1950, 1945, 1940, etc.). Cast Iron accounts for the vast majority of the oldest pipe (pre-1960), hence the average age will be highest. Furthermore, the cast iron pipe also has a longer average monitoring period. This chart shows that the majority of this pre-1960 cast iron pipe was installed by the City of Toronto, which also has the longest running data set at 55 years. The other five utilities average 28.4 years of monitoring. This means that the older cast iron pipes from the City of Toronto will have almost twice as many years of monitoring records as the average among all other pipes.

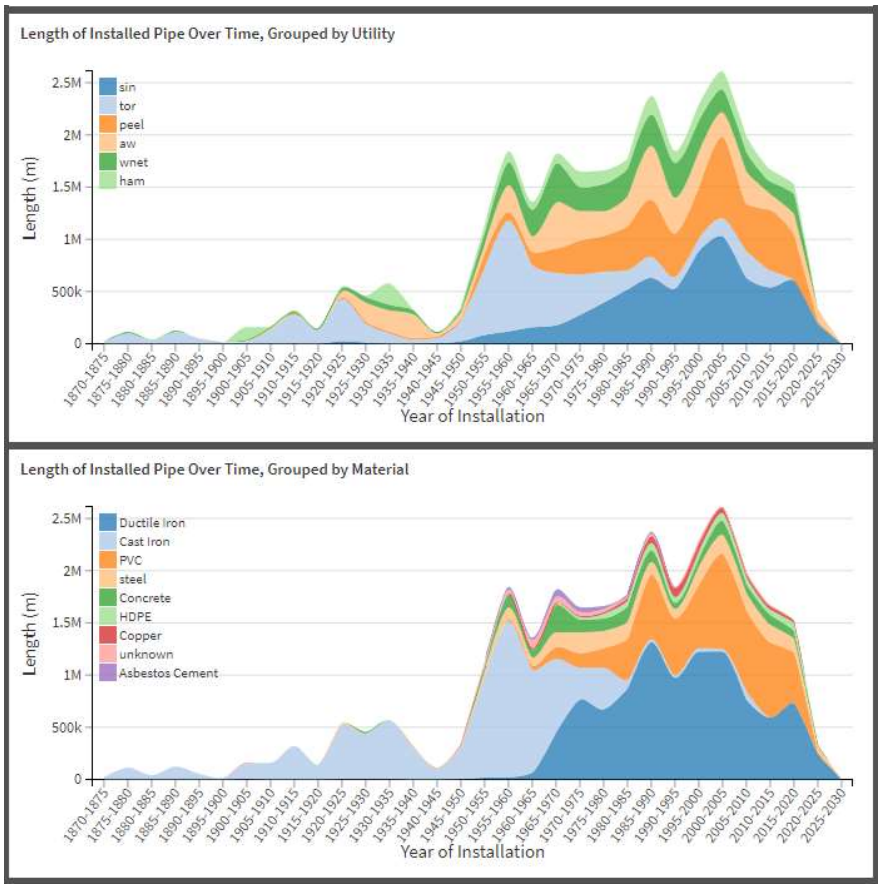


Figure 13: Pipe installation history, grouped by utility (top) and material (bottom).

A breakdown of the number of breaks over time is presented in Figure 14. The number of breaks is heavily impacted by utilities starting and stopping their break recording programs.

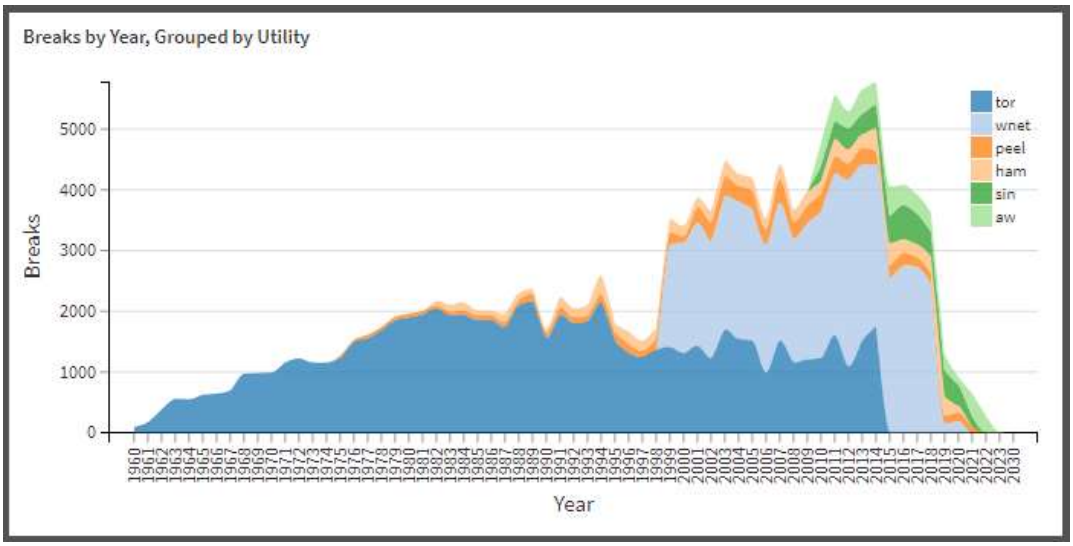


Figure 14: Number of Breaks vs Time, Grouped by Utility.

A crosstab analysis by material and utility is provided in Figure 15. This analysis reveals substantial differences in the break rate (scaled to breaks per 1,000 segments per year), average length, and average age both across materials and utilities.

Break Rate per 1,000 Segments, Age, and Lenght by Utility and Material

		dataset						
mat_bin	Value	aw	ham	peel	sin	tor	wnet	Grand Total
Cast Iron	Average of breaks_per_1000_seg_yr	7.60	12.99	15.34	6.02	37.91	22.29	**19.39**
	Average of age	71.77	81.77	61.76	51.95	76.99	66.05	**68.45**
	Average of len	74.55	60.04	106	33.52	127	28.39	**69.70**
HDPE	Average of breaks_per_1000_seg_yr	0	1.51	2.29	1.73	5.30	19.05	**17.07**
	Average of age	0	21.48	18.72	13.41	14.54	22.01	**21.85**
	Average of len	0	51.07	131	67.67	131	28.32	**32.21**
Copper	Average of breaks_per_1000_seg_yr	0	0.11	1.39	0.59	1.28	24.65	**14.81**
	Average of age	0	26.14	23.05	23.39	34.62	26.38	**25.29**
	Average of len	0	51.19	51.97	17.04	56.64	23.46	**29.98**
Asbestos Cement	Average of breaks_per_1000_seg_yr	2.79	13	11.09	17.63	12.19	13.31	**13.04**
	Average of age	53.98	48	54.65	44.75	45.42	46.97	**47.81**
	Average of len	108	473	110	56.52	285	42.59	**62.80**
PVC	Average of breaks_per_1000_seg_yr	1.87	0.30	0.48	30.90	0.98	15.30	**6.09**
	Average of age	9.05	10.14	17.64	35.73	12.93	31.15	**21.58**
	Average of len	39.46	48.10	88.21	41.49	112	34.16	**66.31**
unknown	Average of breaks_per_1000_seg_yr	0.56	1.66	1.78	4.32	12.90	2.52	**5.31**
	Average of age	30.30	50.59	34.66	49.49	46.58	50.02	**43.36**
	Average of len	37.21	58.05	78.31	50.08	153	20.65	**75.01**
Ductile Iron	Average of breaks_per_1000_seg_yr	1.76	3.02	11.13	1.72	23.99	11.51	**3.79**
	Average of age	25.96	32.10	44.23	20.41	34.54	23.10	**23.54**
	Average of len	63.42	55.55	99.95	32.89	133	21.20	**43.27**
steel	Average of breaks_per_1000_seg_yr	0.21	1.53	1.80	2.14	0.12	1.21	**1.58**
	Average of age	16.80	64.04	36.29	30.79	42.15	43.98	**33.86**
	Average of len	7.20	207	87.87	109	251	24.33	**87.33**
Concrete	Average of breaks_per_1000_seg_yr	0.73	2.24	1.10	0	3.15	0.16	**0.41**
	Average of age	42.77	35.91	30.06	0	39.65	55.64	**50.48**
	Average of len	125	139	155	0	237	11.08	**43.75**
Grand Total	**Average of breaks_per_1000_seg_yr**	**3.49**	**5.35**	**2.94**	**2.48**	**28.16**	**13.39**	**8.23**
	Average of age	**40.57**	**42.61**	**25.58**	**25.93**	**59.58**	**40.32**	**36.05**
	Average of len	**66.41**	**57.50**	**95.12**	**38.83**	**131**	**24.76**	**54.54**

Figure 15: Crosstab analysis by Material and Utility, showing the average breaks per 1,000 segments per year, age, and length for each group.

Perhaps most interesting in this table are the differences in segment length. These differences in length are explained by four primary factors. First, urban areas tend to have shorter segments than rural areas. Second, the definition of a pipe segment is somewhat flexible. As an example, when a customer service line connects to a water main, some utilities may decide that this divides the main into two separate segments, whereas others may choose to call it a single segment. Third, differences in water network construction practices can have significant impacts on the measured pipe lengths.

For example, water distribution mains in North America often use a standard minimum size of 4 inches (or 100 mm) to ensure sufficient flows for fire protection, whereas in Europe it is more common to reduce the size in steps as the end of a branch line is reached. Since each diameter change creates a separate logical segment, this practice results in more segments of shorter lengths. Fourth, data collection differences can have a significant impact. For example, in North America it is common for water meters to be placed relatively close to the road. The customer is responsible for the portion of their service pipe on their side of the meter. The short portion of the service pipe on the utility's side of the meter is often recorded in their systems only as a service connection on the main, rather than as a separate pipe. In Europe and Asia by contrast, water meters are often in the home, and the utility is responsible for the entire service pipe. This would lead to a large number of short, small diameter pipes. As a second example, short (less than one meter) closure pieces of different materials are often needed when constructing a pipeline, which can either be considered part of the segment they are on or as dividing it into three segments (the upstream segment, the closure piece itself, and the downstream segment). The implications of these varying segments lengths are explored in Chapter 2.1.1.1.2.

The variations in break rate by utility, material and other factors will be described in Chapter 4.2. Note that some of the following analysis requires use of the Segment-Year records described in Chapter 3.3.3. While there are over 10,000,000 such rows, due to memory limitations, pre-aggregation was performed across several dimensions of the data to facilitate this analysis.

4.1.3 Confirming Normalization by Pipe Length

If each stick of pipe had the same probability of breaking, then breaks per pipe segment should correlate linearly with segment length. This relationship is apparent in Figure 16 for segments less than 500m long (to the left of the orange lines). The chart on the left presents average break rates for bins of 100m. It shows an obvious trend for lengths below 500m, with the trend breaking down as lengths increase. The chart on the right shows the same data, but with smaller bins of 10 m to confirm the consistency of this relationship for shorter length pipe segments. The left on the right also makes it clear that the trend does not hold for longer segments (to the right of the orange lines).

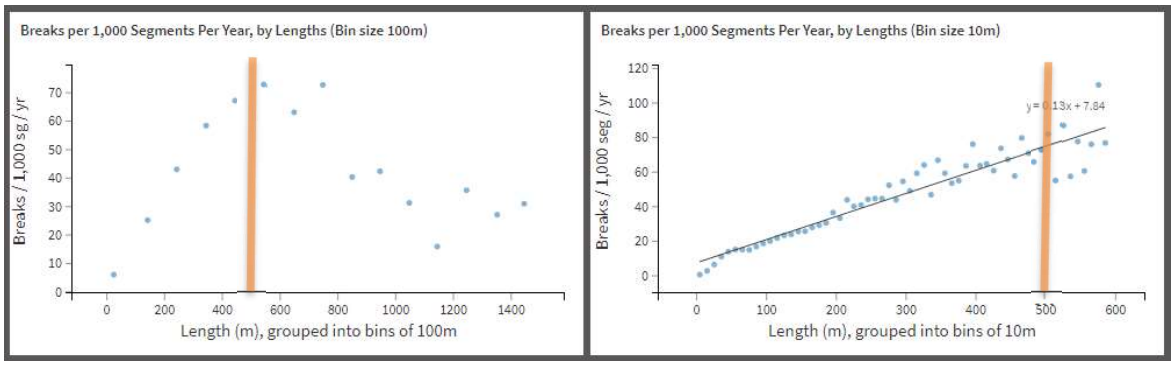


Figure 16: Average breaks per 100 segments per year, among segments of various lengths.

The increasing complexity for longer lengths may be explained by the difference between distribution mains and trunk mains. Distribution mains are the smaller diameter pipes that are available for customer connections. These tend to have segment lengths no longer than the length of a block or the distance between fire hydrants in an urban area (typically 40m to 200m) and will have a consistent design during any era of pipe installation for a given utility. Trunk mains are larger diameter pipes used to bring water to an area of the network. These are often longer, and the selection of material, diameter, and pressure class will be specific to that particular pipe. Figure 17 shows how the material prevalence and average pipe diameter change with segment length.

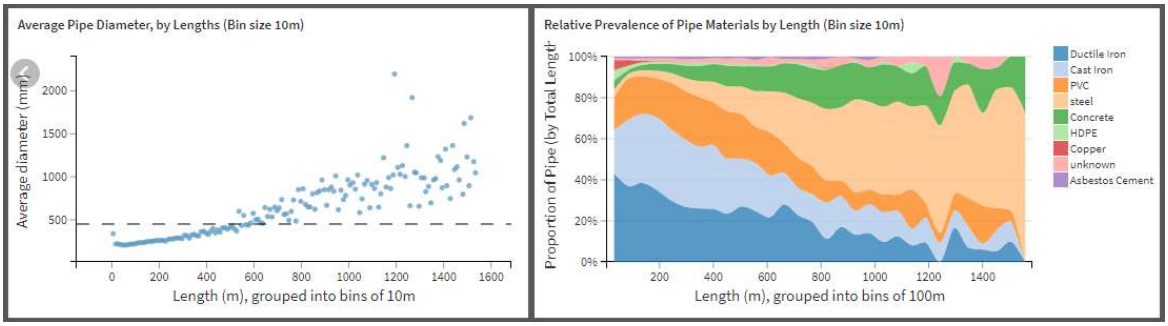


Figure 17: Pipe length (bins of 10m) vs average diameter (left) and relative prevalence of different pipe materials (right).

The linear relationship between length and break rate within more homogenous populations is supported by more detailed analysis. Figure 18 shows this relationship for Cast Iron pipe only, with both the less granular (bin size of 100m) and more granular (bin size of 10m) charts showing a strong linear relationship with nearly identical slope. Similarly, the length and break rate relationship is

shown for Ductile Iron pipe only in Figure 19, which also shows linear relationships in each plot with nearly identical slopes.

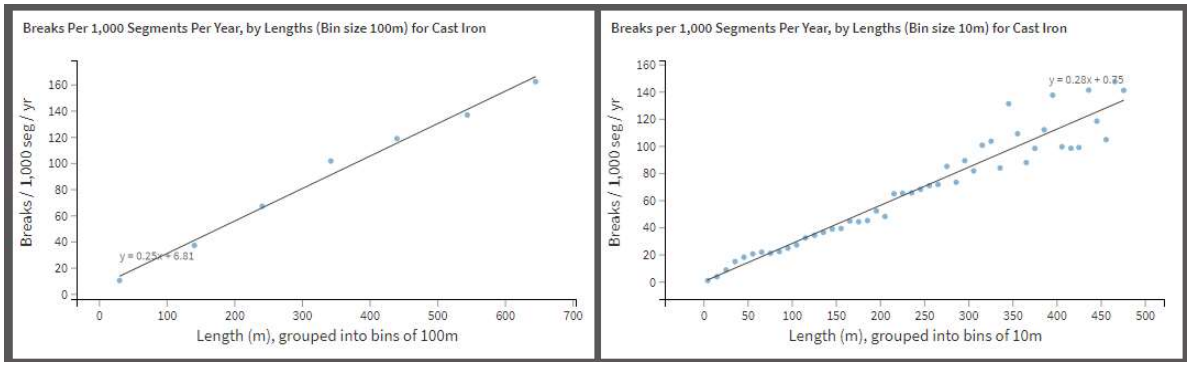


Figure 18: Average segment break rates for Cast Iron, with bin sizes of 100m (left) and 10m (right), in both cases only showing bins with at least 2,500 segment-years of data.

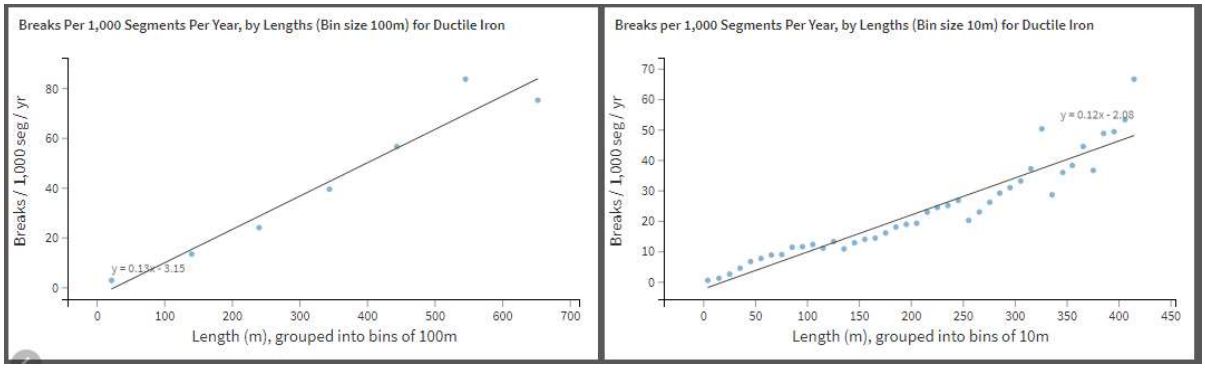


Figure 19: Average segment break rates for Ductile Iron, with bin sizes of 100m (left) and 10m (right), in both cases only showing bins with at least 2,500 segment-years of data.

The linear relationship between segment length and break rate within a homogenous group of pipe types is further confirmed by isolating particular material-diameter combinations. Figure 20 shows the break rate per 1,000 segments per year for various groupings of diameter (bins of size 100mm) and length (bins if size 10m). The diameter bins are labeled by their midpoint, with pipes of diameters in exact multiples of 100 assigned upwards (i.e., the 150mm bin includes pipes with diameters from 100mm to 199.99mm, the 250mm bin includes pipes with diameters from 200mm to 299.99mm, etc.). In each case, only groups with at least 1,000 segment-years included are shown in the charts. The top chart including all four categories shows visually distinct trends for each of the four, with the clear pattern of larger diameters failing less often at a given length. Each of the four

lower charts shows a clear linear relationship between segment length and the failure rate, with increasingly noisy results in the larger diameters and lengths likely due to smaller sample sizes.

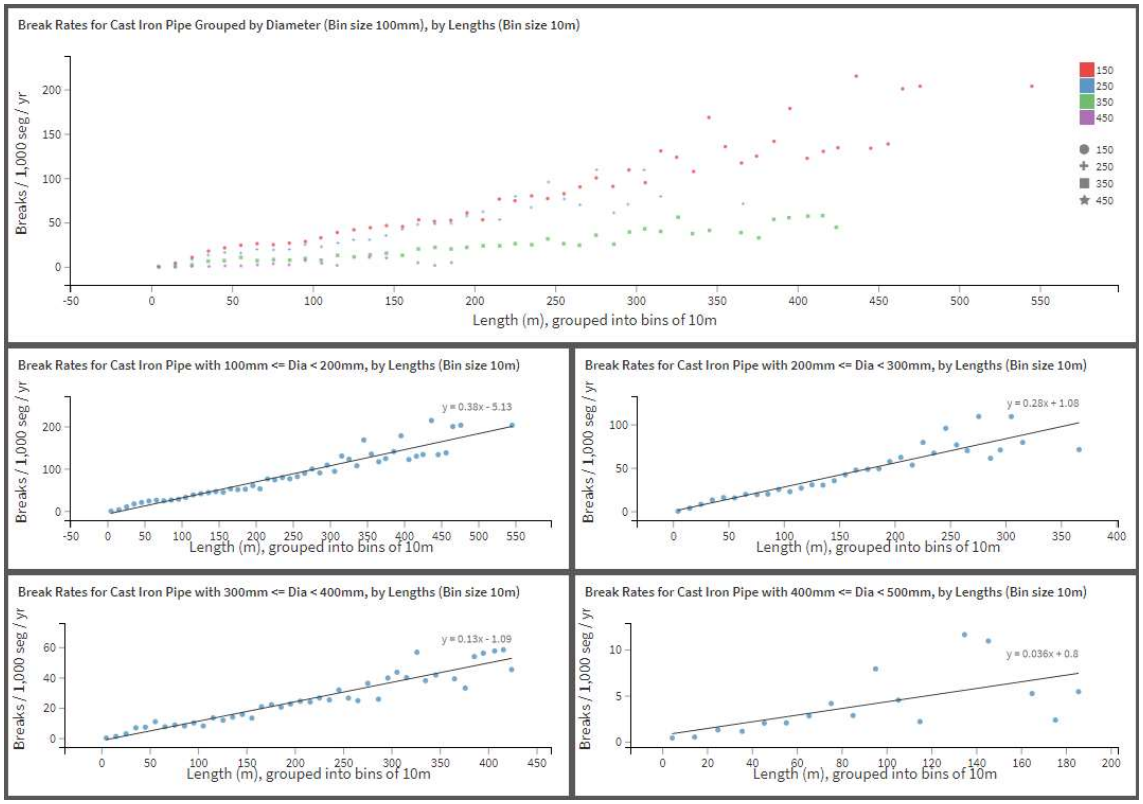


Figure 20: Average segment length vs break rate for Cast Iron (CI) pipe of four different diameter groupings, shown together (top chart) and individually (bottom four charts).

Once again, the trend in Ductile Iron confirms the finding within Cast Iron, as shown in Figure 21. The four trends in the top chart are less visually distinct; however, once again, the charts of individual material-diameter combinations show clear linear relationships, again increasingly noisy for larger diameters and longer lengths.

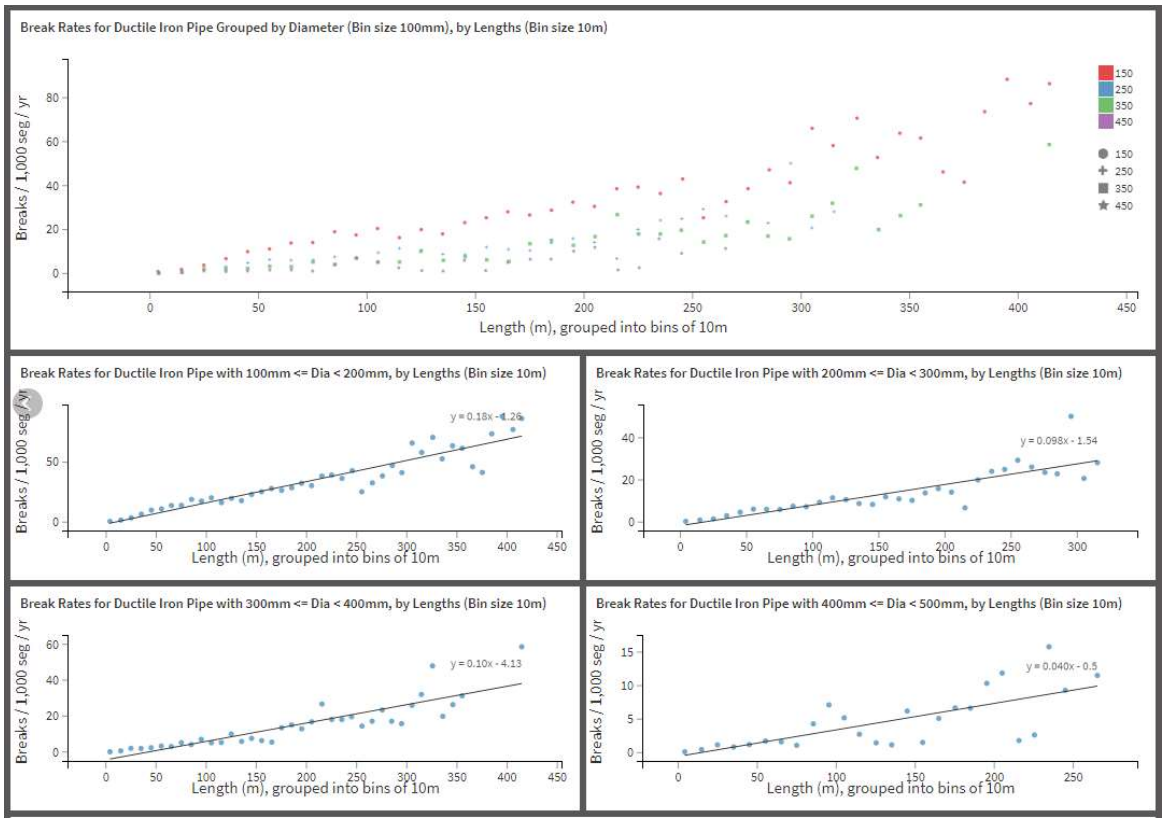


Figure 21: Average segment length vs break rate for Ductile Iron (DI) pipe of four different diameter groupings, shown together (top chart) and individually (bottom four charts).

The linear relationship between length and break rate per 1,000 segments per year for a homogenous group of pipes thus supports the intuitive hypothesis that break rates can reasonably be normalized by segment length. This is highly convenient when considering groupings or cohorts of pipes, where normalizing the break rate per 100 km per year has become common in the literature. In keeping with this common standard, the predictive feature analysis performed in Chapter 4.2 has used break rates per 100 km per year.

4.2 Individual Predictive Feature Analysis

This section provides a brief view of individual predictive factors. The data broadly conforms to the prior literature, with a few surprises and novel observations. The predictive feature analysis was conducted towards the goal of effective feature engineering for the predictive models. While certain

observed relationships may warrant a more formal statistical and causal analysis, this would be outside the scope of this thesis and is left as an area for future research.

4.2.1 Pipe Diameter

Figure 22 shows the expected relationship between diameter and break rate using both breaks per 1,000 segment-years and breaks per 100 km-years. When the break rate is plotted on the basis of km-years, the relationship appears visually similar to an inverse relationship.

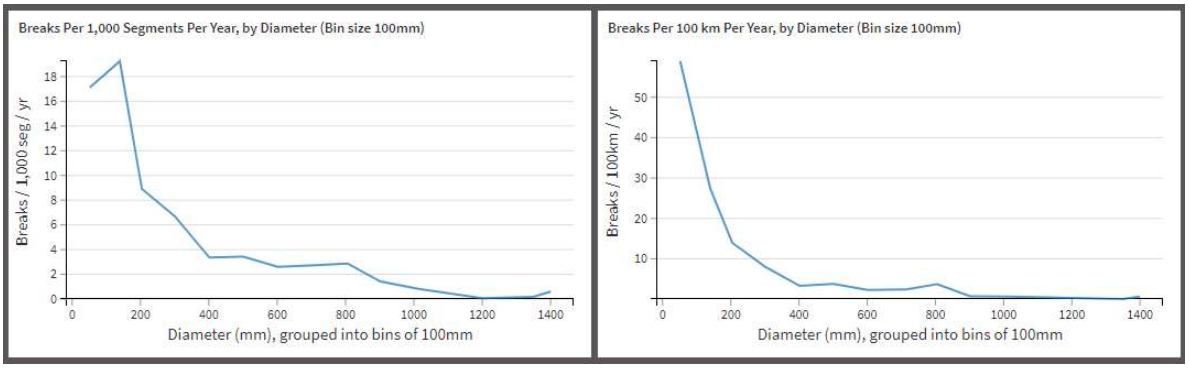


Figure 22: Break rates per 1,000 segment-years (left) and per 100 km-years (right), grouped by diameter with bin sizes of 100mm.

It is immediately apparent that diameter is an extremely strong predictor of failure rate, with smaller diameters being more likely to fail. In the per 1,000 segment-years plot, the dip in the smallest diameter category (<100mm) is likely due to a large number of fittings (valves, etc.) and service connections of extremely short length, which appear only in these very small diameters. This relationship is confirmed by the per 100 km-years plot, where the same diameter category (<100mm) has the highest failure rate. The average failure rate of small diameter pipe (<100mm, covering 1.02 million km-years of data) is 116.5 breaks / 100 km / yr, whereas the failure for large diameter pipe (>=900mm) are 0.74 breaks / 100 km / yr or lower in each of the 16 size bins (covering 1.95 million km-years of data in total among the 16 bins). This ratio of over 100:1 in the break rate offers a powerful predictive tool. As shown in Figure 23, the trend holds with a finer-grained grouping into diameter bins of 25mm (left) as well as when the analysis is extended out to very large diameters (right).

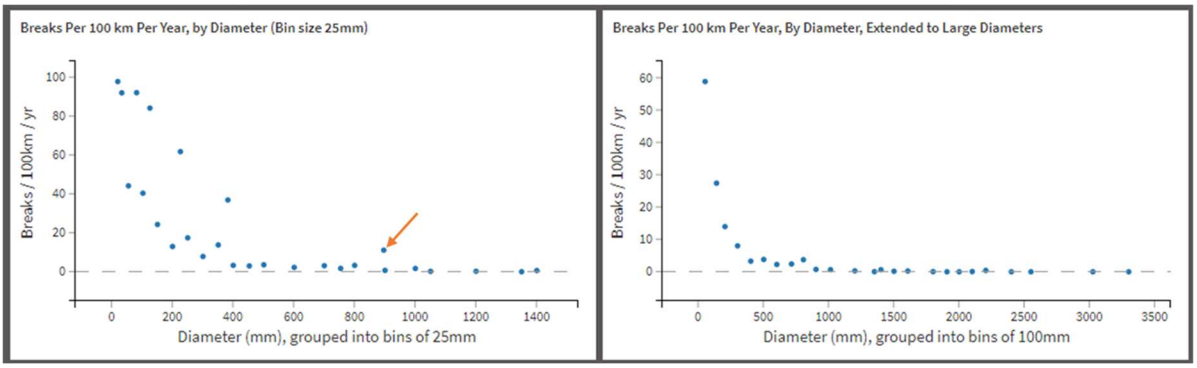


Figure 23: Break rates per 100 km-years, with a finer-grained presentation at smaller diameters (left) grouped by diameter with bin sizes of 25mm, as well as an extension up to large diameters (right) grouped by diameter with bin sizes of 100mm.

The data is increasingly noisy with the 25mm bin sizes, which is a consequence of different pipe diameter standards being used in different jurisdictions. While all utilities use relatively similar pipe diameters, there are minor variations. In a coarse binning scheme with bins of 100mm, regardless of these variations, each utility will have at least one standard diameter falling into each bin. However, in the fine-grained binning scheme with bins of 25mm, several bins contain data from just a single utility. In such instances, a particular data quality issue present in a single utility will not be averaged out by other utilities. In these cases, a single data quality issue can lead to an outlier in the data. An example is the outlier in Figure 23 in the diameter bin of 875mm to 900mm, with average diameter of 896mm, as highlighted with the orange arrow. This bin shows a break rate of 11.1 breaks / 100km / year and contains a total of 31 breaks. Of these, 29 are attributed to a single pipe segment 17.16m in length. It seems implausible for such a short segment to have suffered so many breaks, and investigation confirmed this to be a data quality issue. The address records for these breaks are all along the same street, with street numbers ranging from 1132 through 4831. However, manual investigation of these addresses via Google Maps shows that no addresses numbered above 1100 on this street are found. They instead return a location at the midpoint along this street. That point is located along this short length of pipe. With these 29 erroneous leaks removed, the break rate for this entire bin declines to 0.7 breaks / 100km / year, which is well aligned with the neighboring points and overall trend. This single example serves to illustrate the point that some degree of noise is expected in these relationships, due to the unavoidable data quality issues present in such a large manually compiled dataset.

The trend is clearly present in five of the seven common pipe materials: Cast Iron, Ductile Iron, HDPE, PVC, and Asbestos Cement. As seen in Figure 24, all the diameter lines are substantially higher in the small diameters than the large diameters. Small diameter steel is the exception, with a low break rate among very small diameter pipes. Concrete is omitted from the analysis, as small diameter concrete pipe is not often used for water mains, and none of the utilities in this study reported small diameter concrete pipe in use.

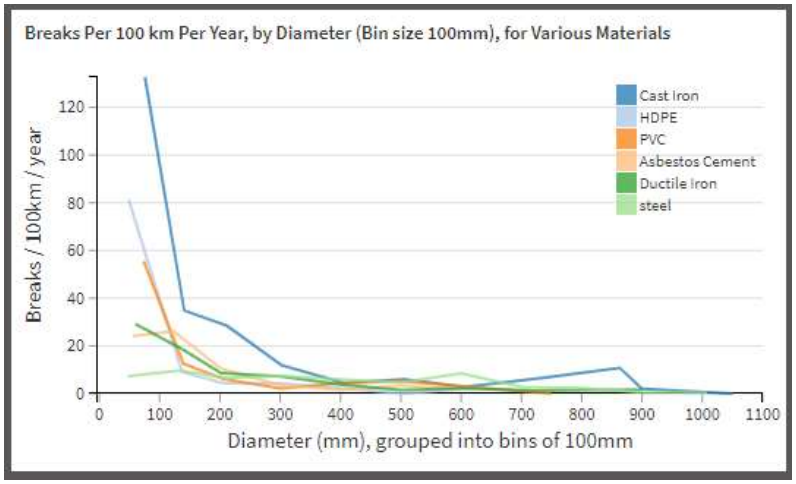


Figure 24: Break rate per 100 km per year, grouped by diameter with bin sizes of 100mm, for various materials.

Deeper investigation into this relationship for steel pipes shows that the trend is generally present, as shown in Figure 25, except for two clear outliers. There is a lower-than-expected break rate among very small diameter (<100mm) steel mains. This data point contains a relatively small amount of data (4.6 km of pipe), the majority of which (3.8 km) comes from one utility (American Water). Manual analysis suggests this outlier represents a data completeness issue. The American Water breaks data included a record of the diameter and material measured when the break was repaired. A query was applied for pipes with a measured diameter of 100mm or less, where the measured material did not match the material for the pipe to which the break was associated. This revealed 17 instances of a pipe which was measured as steel, but the pipe record indicated another material, all of which were short customer service lines. There were only two instances of the pipe record showing steel and the measurement at the site of the break showing steel. Reversing these misallocations would add another 15 breaks to the three currently tracked in this bin. That would raise the break rate from 7.3 to 43.8, putting it well in line with the trend. A hint as to the reason for this asymmetry of errors

comes from examining the pipe installation and break dates. For the majority of these 17 instances, the break occurred before the installation date listed for the pipe. This suggests that the failure occurred on an older steel main, which was subsequently replaced by a new pipe of a different material (generally PVC). This asymmetric replacement pattern would result in exactly the asymmetric material errors found in the data. The second outlier showing an unexpectedly high break rate is for the 600mm to 699mm diameter bin. Of the 55 breaks in this bin, at least 23 are due to data quality issues. These 23 leaks are registered to a single line just 25 meters in length. The recorded addresses all failed to include a house number, so were geocoded to the beginning of the street, where this pipe happened to be. All but three of the breaks in this bin came from the same utility. Manual inspection of the break records from an additional three pipes accounting for a further 16 breaks showed seven more with blank house numbers and eight with addresses registered to houseboats with comments suggesting that the issue was in fact with a customer service connection rather than the adjacent large diameter steel watermain. Accounting for these data quality issues, the steel pipe category would join the remaining five materials analyzed as having a clear trend whereby the break rate decreases as diameter increases.

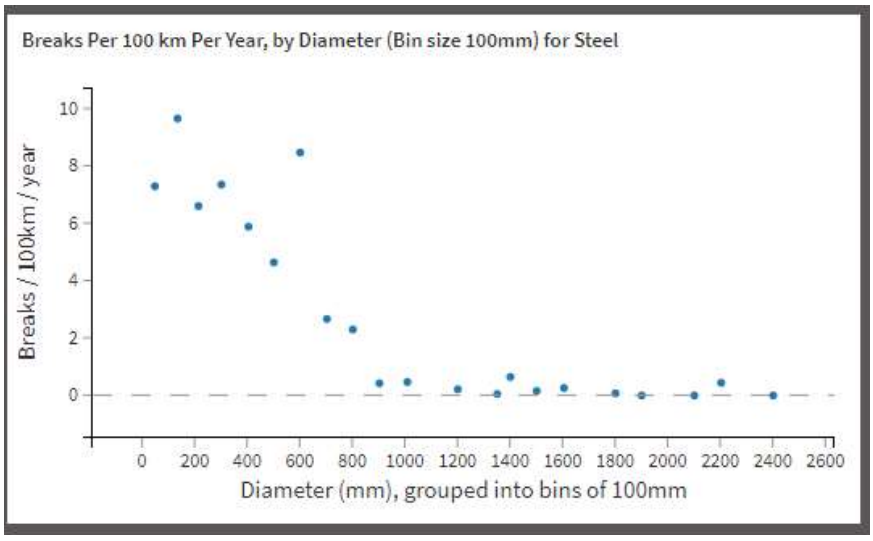


Figure 25: Break rate per 100 km per year, grouped by diameter with bin sizes of 100mm, for steel.

Despite the noise introduced by data quality issues, the relationship of the break rate increasing as pipe diameter decreases remains strong, and visually approximates an inverse relationship. This apparent inverse relationship between diameter and break rate was confirmed by inverting each of the

x and y axis variables. As shown in Figure 26, the resulting relationship appears linear, regardless of whether the inverse of average diameter is used for the x-axis, or the inverse of the break rate is used for the y-axis.

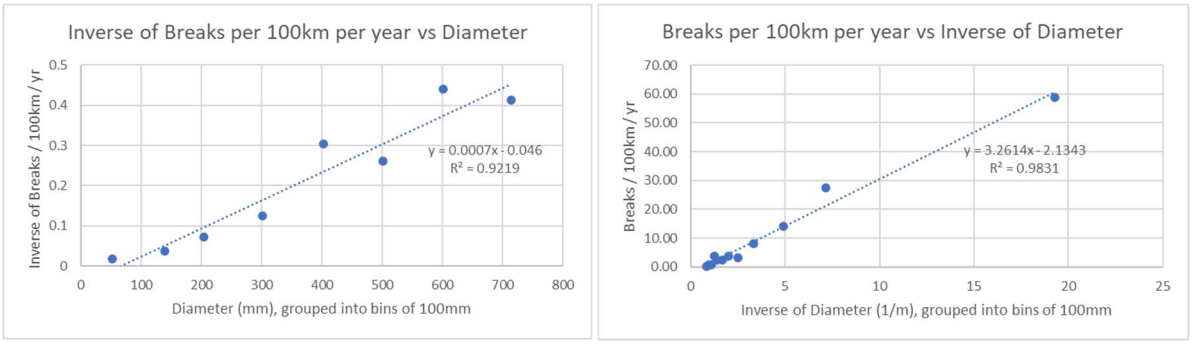


Figure 26: Break rate vs diameter, with inverse variables for the y-axis (left) and x-axis (right), confirms the inverse linear relationship between diameter and break rate.

The linear relationship holds true when tested for individual pipe materials, as shown in Figure 27. With the increasingly fine-grained analysis, some of the categories contained relatively small numbers of breaks and km-years of pipe history. Thresholds of five breaks and 10 km of pipe were established to limit this effect; nevertheless, the data is notably noisier.

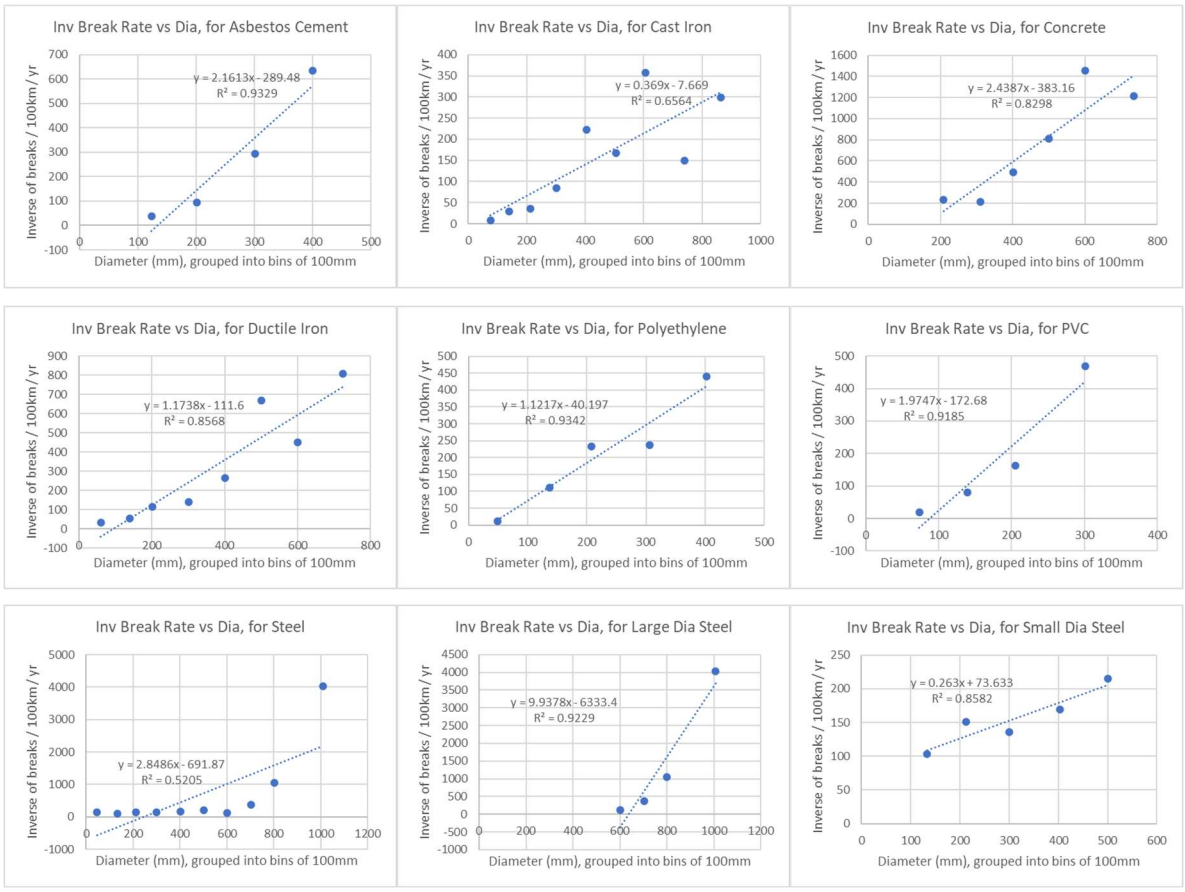


Figure 27: Inverse of break rate per 100 km per year, grouped by diameter with bin size of 100 mm, for each of the seven materials with multiple bins containing at least five breaks and 10 km of pipe.

Once again, some of the noise in these relationships has been confirmed as due to data quality issues. An example of this comes from the Concrete mains of 800mm diameter. This is an outlier from the trend line, with an inverse break rate much lower than the trend (i.e., break rate much higher than the trend). This group contains a total of 41 break records. However, all these breaks come from the same utility, and at least 34 of the 41 breaks have been confirmed to be data quality issues from the geospatial mapping step. These 34 breaks were tracked to two pipe segments, one of which had a length of 65m and eight breaks, and the other a length of 436 m with 26 breaks. For the pipe with eight breaks, examination of the comments showed that all were on customer connections attached to a smaller 42mm diameter copper distribution main running underneath the street. The 800mm concrete main was in a right of way through the front yards of the houses. The breaks were

geospatially mapped to the closest main but were in fact on a different one. This illustrates a potential for improving the geospatial mapping step, with such work being outside the scope of this study. The 26 breaks were all due to an error in the geocoding service used. These breaks were all recorded at addresses of houseboats in Amsterdam. The geocoding service used is not able to place houseboat locations, so instead mapped all 26 to the lowest house number on their street. This incorrect geocoding resulted in them being matched by proximity to the wrong pipe. Removing these errors would change its inverse break rate to 0.7, placing this data point much closer to the trend line.

While these data quality issues represent a tiny percentage of the overall dataset, due to the small lengths of pipe and numbers of breaks in certain material-diameter bins, they can introduce significant noise to individual data points in this fine-grained analysis. This sort of noise is expected for the machine learning modeling that is the primary focus of this study, which should perform well despite noisy data. It is called out in this section to note the trend that wherever an outlier point from the trend was investigated, a data quality issue was discovered, whereas no data quality issues could be identified in a selection of data points investigated along the trend line.

4.2.2 Pipe Material

Break rates by material, as shown in Figure 28, were largely as expected from the literature review. The least break-prone materials by all three normalization methods are steel and concrete. Cast Iron and HDPE are among the most break prone, regardless of normalization method. Copper is used exclusively in small diameter, short length service connections, so shows a very high break rate per 100km, but performs nearly identically to PVC (a newer material used in a similar manner) when normalized for diameter.

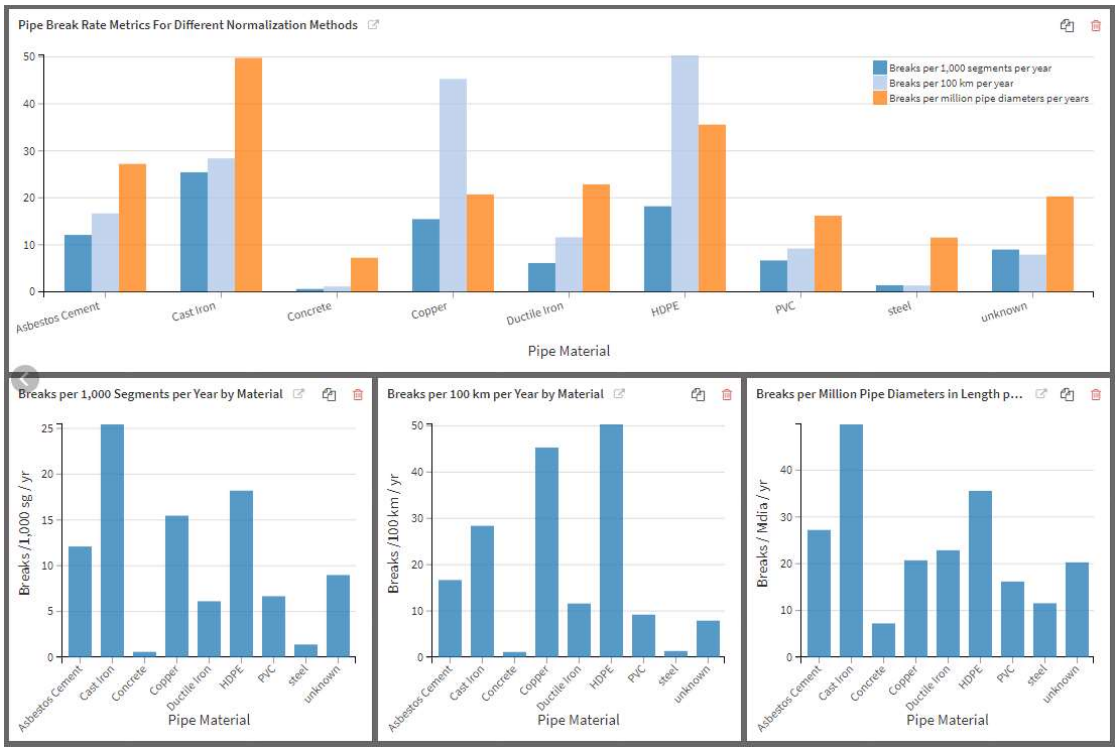


Figure 28: Break rates by material length and break counts, respectively.

It is worth noting the apparent effectiveness of the pipe diameter normalization introduced in Chapter 4.2.1. The relative difference between material performance becomes much smaller under this normalization scheme. In particular, the advantage of steel and concrete over other materials becomes much less significant once this diameter normalization is applied.

4.2.3 Utility

It was anticipated that there would be significant variations in the break rate from one utility to another. There are three significant reasons why this variation was expected.

First is administrative differences in the recordkeeping from one utility to the next. The rigor and consistency of recording pipe breaks may vary from one utility to the next, as well as within one utility over the years or decades of the data provided by that utility. There is also no universal consensus as to what does and does not constitute a pipe break. A significant rupture of the pipe body itself, such as a longitudinal split (also called a burst), a circumferential split (also called a "broken back"), or a corrosion blowout, can be expected to qualify for any utility. On the other hand, a small leak, proactively detected and repaired, may or may not be tracked as a break depending on the

utility's practices. This is even more true if the leak occurs on an appurtenance, such as a valve, fire hydrant, or meter connected to the pipe, rather than on the pipe itself. Customer service connections offer perhaps the most substantial opportunity for administrative differences. These are the short (typically under 3m to 10m) lines leading from the water main to the individual property, with a water meter set in between. Typically, the utility is responsible for the connection from the main up to the water meter, and the customer is responsible for the connection from the water meter to the home. These connections may be tracked by a utility as separate pipes or may be tracked simply as customer connection points along the main pipe. Breaks on these service connections are quite common and may be recorded to those small pipes themselves, to the main pipe that they attach to, or not recorded at all.

Second is differences in the pipe population. The age, diameter, and material of pipes may vary substantially from one utility to another. These differences should not impact the analysis by other parameters (such as diameter, material, and age), nor the performance of a machine learning model.

Third is the impact of the local environment. This includes the soil environment (corrosivity, and stability), the water inside the mains (water hardness and presence of pressure transients), the climate and weather, and finally the commercial environment (pipe manufacture and installation practices).

While the data presented cannot distinguish between these causes, there are clearly differences between the participating utilities. Figure 29 shows the break rates by utility and material. One utility (Waternet) clearly stands out with the highest break rate in all materials except for PVC, in some cases more than three times as high as the next highest utility. The same trend holds in Figure 30, which provides a breakdown by diameter bin and utility, where Waternet has the highest break rate in every diameter bin except one. As Waternet enjoys moderate climates, stable soils, non-corrosive water, and world-class engineering and construction standards, these differences can be explained only by differences in administrative processes. This observation is backed up by manual inspection of the break records. Upon translation from Dutch, these reveal repairs to valves and water meters, reports of leaks inside customer properties, and a wide range of other issues which most utilities would not track as a leak, break, or defect.

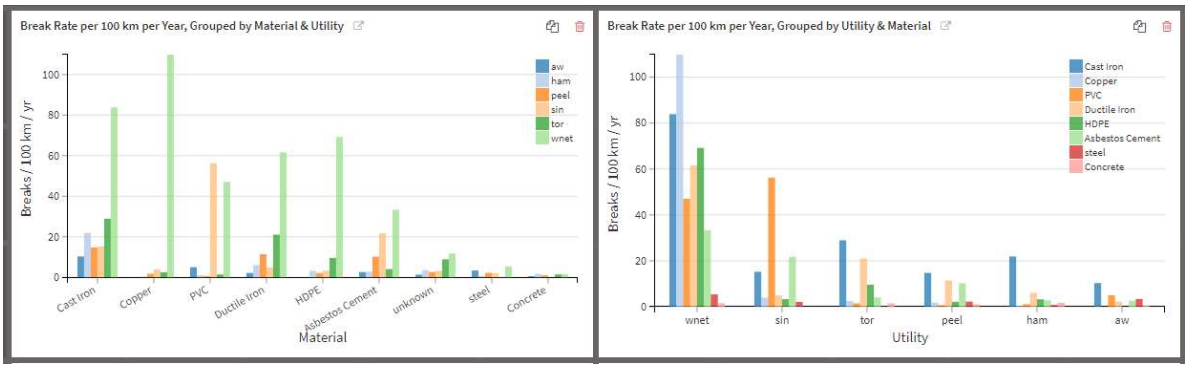


Figure 29: Break rates per 100 km per year, grouped by utility and material.

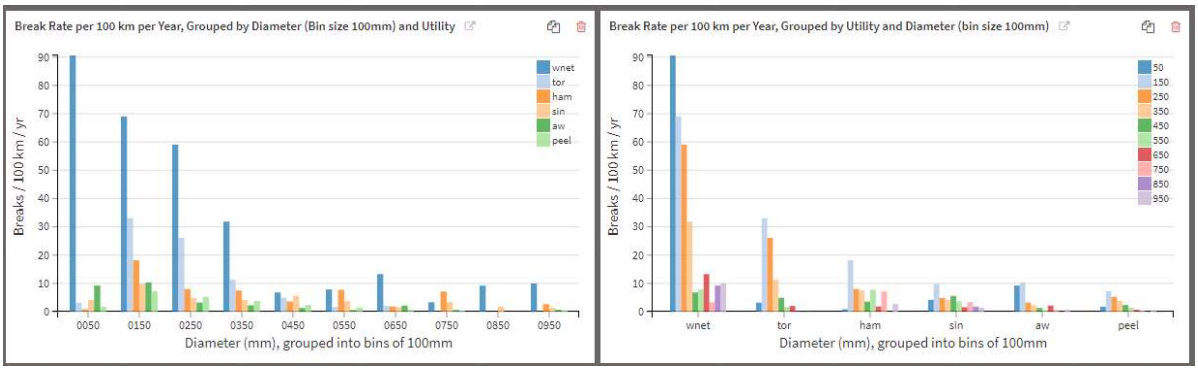


Figure 30: Break rates per 100 km per year, grouped by utility and diameter.

The remaining utilities show comparable break rates when broken down by material and diameter. There is no clear standout, with each utility showing low break rates in certain materials and diameters, and higher break rates in others. This result is as expected, as all the participating utilities are known as global leaders in practices and technology for managing their water networks. One example is Singapore PUB, which shows the lowest break rate per segment, and below average per-km break rates in every diameter category. Singapore exists in a warm climate, and high temperatures are well known to increase the pace of corrosion, which is a temperature-dependent electrochemical process. It also sits within the "ring of fire" region prone to earthquakes and earth movement. Singapore PUB is world renowned for the level of investment and care it takes of its water network, preferring quality over low cost for its pipe materials and implementing one of the most rigorous leak detection programs in the world. These commercial factors likely counterbalance the challenging environmental factors faced by this utility. This variety of performance across

diameters and materials does, however, suggest that a wide range of administrative practices are employed for tracking breaks and leaks.

There is no objectively right or wrong way to perform these administrative processes. As such, this variation across utilities is a positive outcome for this study. It means that the dataset represents a diverse range of practices from around the world, and that a model which performs effectively across all these utilities is likely to extrapolate well to new utilities.

4.2.4 Pipe Age

Perhaps most interesting is the relationship between break rate and age. The intuition that older pipes should fail more often is so powerful that the actual relationship has scarcely been studied. The data provided for this study, however, shows that the relationship is far more complex and nuanced than expected.

Prior to an effective analysis of the impact of pipe age, data cleansing activities were needed. The records provided contain numerous instances of pipe installation occurring prior to the advent of the pipe material in question. Filtering of records was performed per the following estimated earliest dates of manufacture for certain relatively new pipe materials:

- Ductile Iron: 1948 (Gray et al., 2009)
- Asbestos Cement: 1929 (Hu et al., 2008)
- PVC in Europe: 1934 (Hülsmann & Nowack, 2004)
- PVC in North America: 1950 (Walker, 2011)
- Concrete: 1910 (Erdogmus et al., 2010)

Immediately apparent upon review of Figure 31 is a replication of the results found in (Sundahl, 1996) in their study of cast iron pipes: there is a direct relationship between age and failure rate for the first 30 to 40 years of pipe life, after which the relationship with age changes. This remains true whether the break rate is unnormalized (breaks per 1,000 pipe segments), or normalized by distance (per 100 km), or by distance and pipe diameter (per million pipe diameters of pipe length).

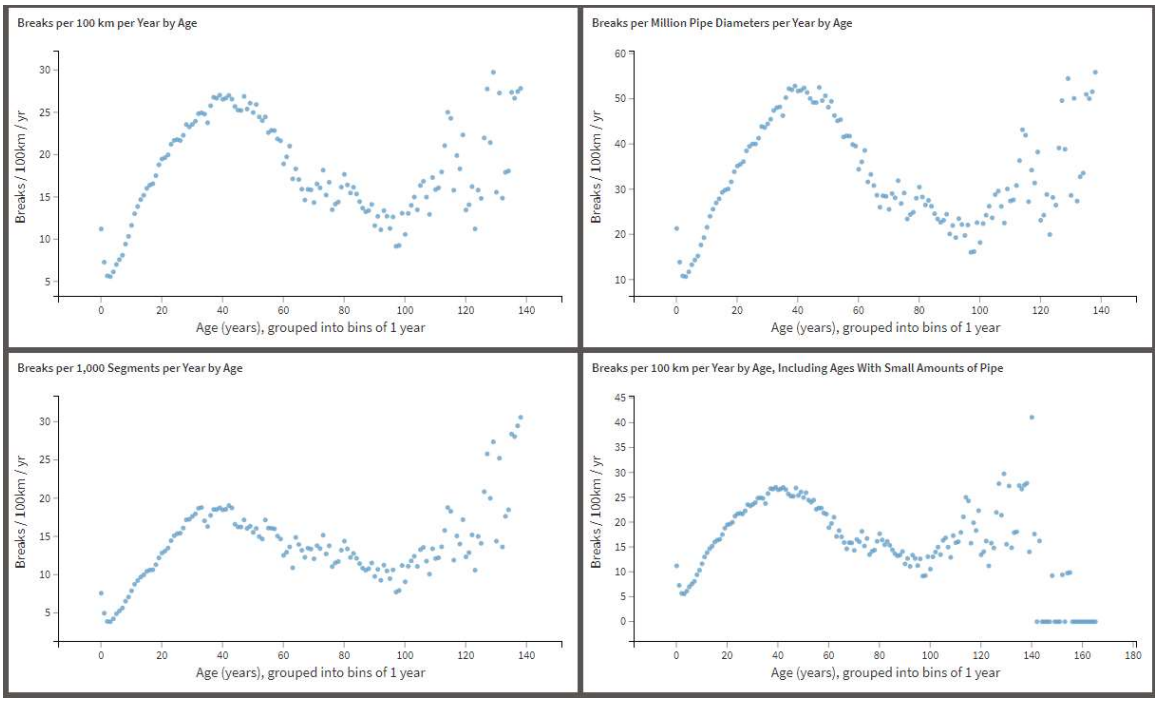


Figure 31: Break rates by age for various normalization methods; all charts except the lower-right were filtered with a minimum threshold of 50km of pipe.

Two additional features of the age / break rate curve are also visually apparent. First, close inspection of the first 10 years reveal the expected bathtub curve pattern (Singh & Adachi, 2013). Break rates begin moderately high, decline for the first two to three years, and then begin to climb. Beyond 40 years, however, an unexpected result appears. Rather than continuing to climb, the break rates peak at around 40 to 45 years, and then begin to decline. The decline is quite pronounced in both of the normalized views and continues until around the 100-year mark, at which point the rate begins another climb. The data beyond the 100-year mark is sparse, hence the relationship between age and break rate from 100 to 140 years is quite noisy. A minimum threshold of 50km of pipe was applied to each 1-year bin of pipe age to all charts except the bottom-right one. This chart shows another sharp decline in the break rate after 140 years; however, the abruptness of the drop-off and the small sample sizes at the ages make this additional decline quite suspect.

This peak-and-trough pattern was first noted during a preliminary investigation of data from Toronto Water (the first utility to provide data). Figure 32 shows a breakdown of data from this utility only, presenting both the break rates for each of the top materials by year and by pipe age. The

data is quite noisy; however, after applying Hamming smoothing (11 point window for year, 31 point for age), consistent patterns begin to emerge. In the lower right chart of Figure 32, each pipe material shows an initial wave of failures, with the peak falling at between 10 and 50 years of age, as illustrated by the star symbol on the chart. The failure rate then declines for a period of time and then rises again later on. This may be evidence of the bathtub curve concept in action, but with the "wear-in" period taking decades rather than the generally assumed few years. The exception is Ductile Iron pipe, which shows multiple waves. This may be due to changes in the manufacturing standards, as early Ductile Iron pipes lacked adequate corrosion protection. This may also be an artifact of data quality issues, as the third and fourth waves of Ductile Iron failures correspond to installation dates prior to the invention of Ductile Iron.

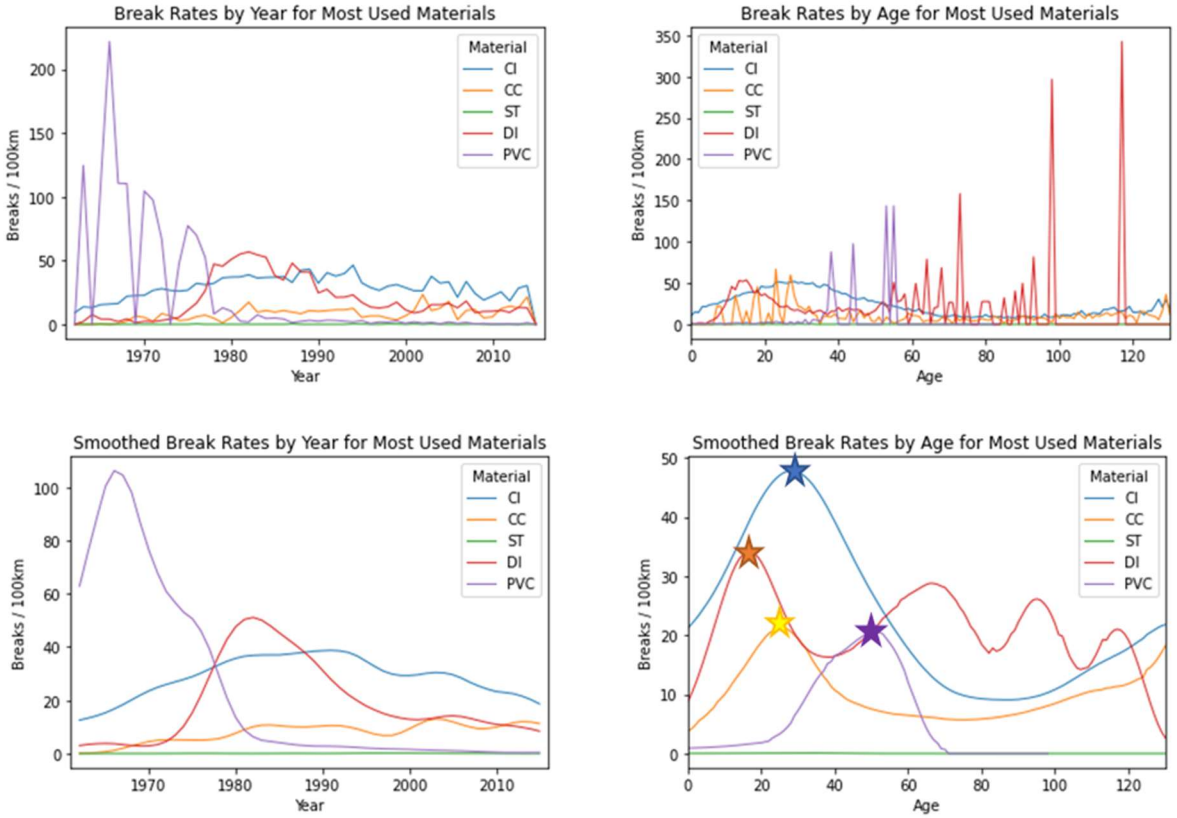


Figure 32: Break rates for Toronto Water data only, by year and age, raw (above) and smoothed (below) with 11 and 31 point Hamming windows for Year and Age respectively.

Extending the analysis to the full data set from all utilities shows that the pattern above continues to hold. To accommodate different scales for the y-axis, materials with moderate to high break rates are

shown in Figure 33, and for low break rates in Figure 34. These charts have the pipes grouped into age bins of 10 years to provide smoothing and have a minimum threshold of 10 km-years (1 km per year of bin size) applied to filter out bins with small sample sizes. Each material follows the general pattern of starting low, rising to a peak, and then falling again. Some materials (such as Copper most clearly) show an additional subsequent rise, and some show another fall after that rise.

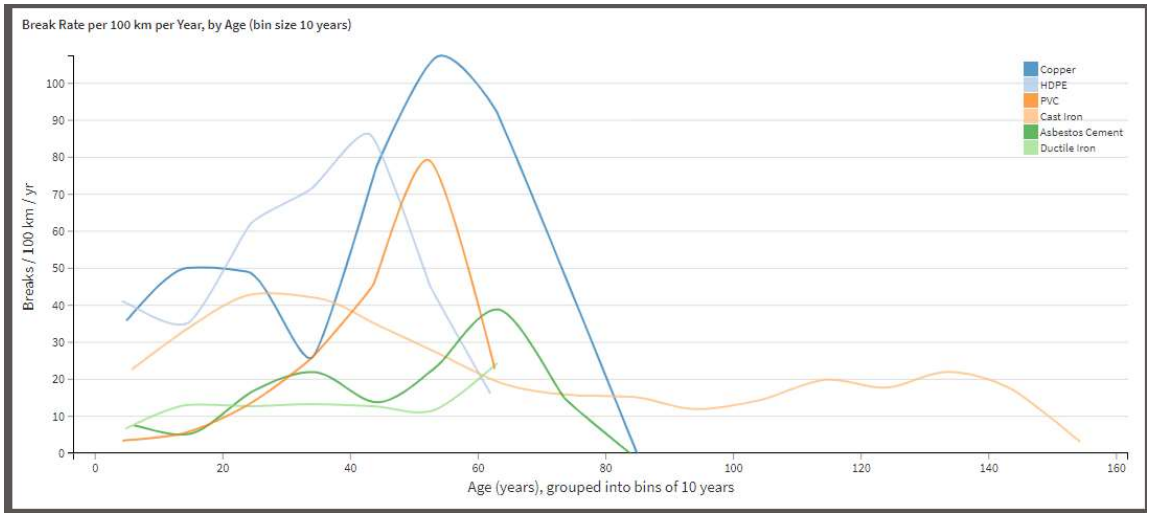


Figure 33: Break rates per 100 km per year, grouped by age with bin sizes of 10 years, for materials with moderate to high break rates.

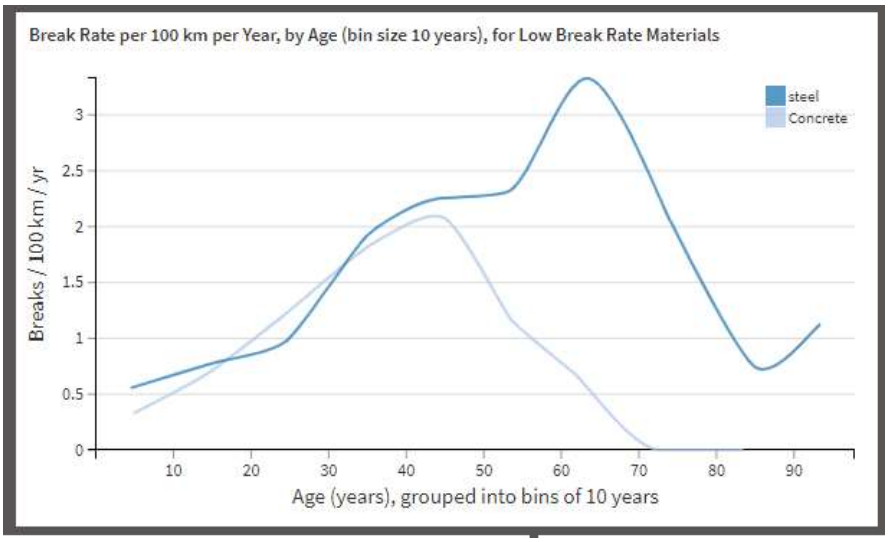


Figure 34: Break rates per 100 km per year, grouped by age with bin sizes of 10 years, for materials with low break rates.

Individual charts for each material are shown in Figure 35. These present highly granular data, with one point on the chart per year of age, using a minimum threshold of 1km of pipe to include an age bin in the chart. For most materials (Cast Iron, Copper, HDPE, Asbestos Cement, and Concrete), the initial rise and fall remains very clear. The pattern is also present for ductile iron, but the peak is low and broad, and the fall is less pronounced. For Steel and PVC, the rise is clear; however, it is unclear whether the subsequent fall is genuine or an artifact of sparse data in the high-age range for those materials.

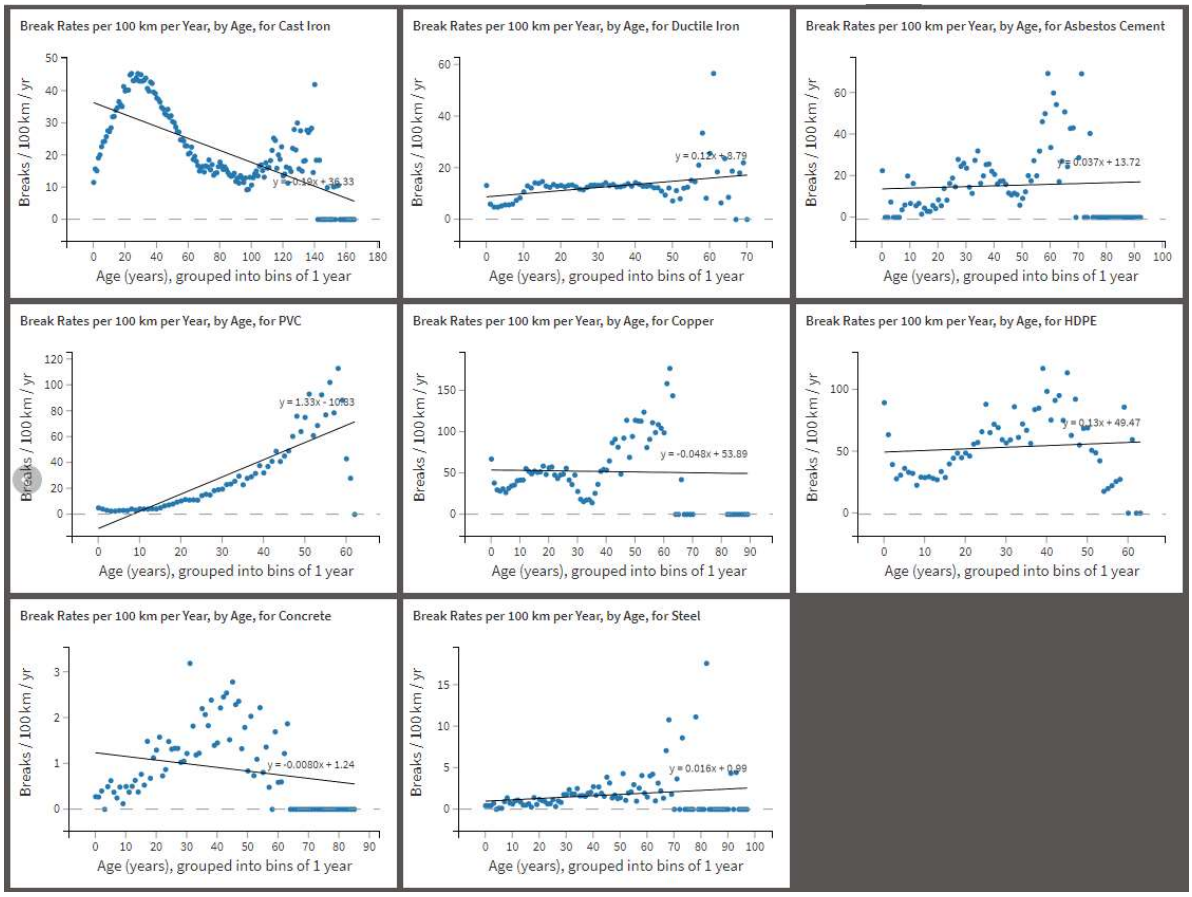


Figure 35: Break rates per 100 km per year, by age, for each material type.

The best fit linear regression lines provided with each chart in Figure 35 illustrate the complexity of the age relationship. For some materials (such as PVC), the line shows a reasonable fit with a slope in the expected direction of increasing break rates with age. For others (such as Cast Iron), the fit appears somewhat reasonable but the slope is in the opposite direction, with failure rates decreasing with age. For others, such as Copper, the regression line appears to offer no predictive value. While

age does appear to be a predictor of failure rate within a given material type, the relationship is certainly not linear or even monotonic.

Proceeding to an even finer grained division of the records, Figure 36 and Figure 37 show the age – break rate relationship for each material-utility pairing. The number of records per year becomes quite small at this level of division, creating highly noisy charts. Binning is applied into bins of five years (for Cast Iron, Ductile Iron, and Asbestos Cement) or 10 years (all other materials) to make the charts more readable. In Figure 37 the charts for Copper and HDPE are presented both with and without Waternet, which exhibits break rates for these materials on a wholly different scale from the other utilities. This difference is likely due to data recording practices, as manual inspection of the break notes suggests that Waternet alone includes breaks on the service connections themselves (including the meters) in their records. Since copper and HDPE are used almost exclusively for these service connections, this would explain the difference.

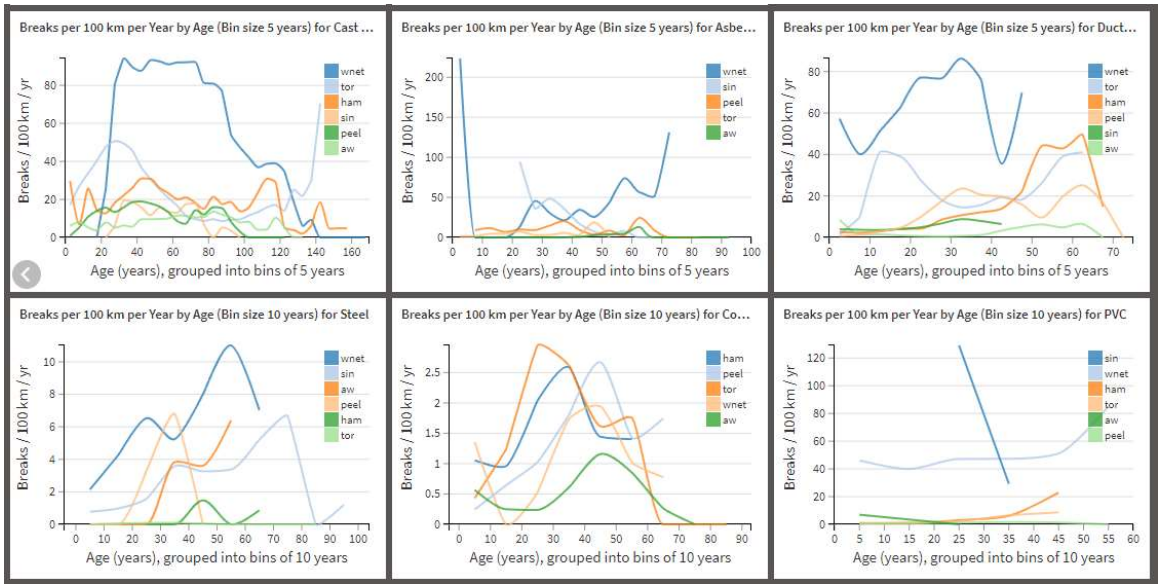


Figure 36: Break rates by age per utility, grouped into bins of 5 years (top) or 10 years (bottom).

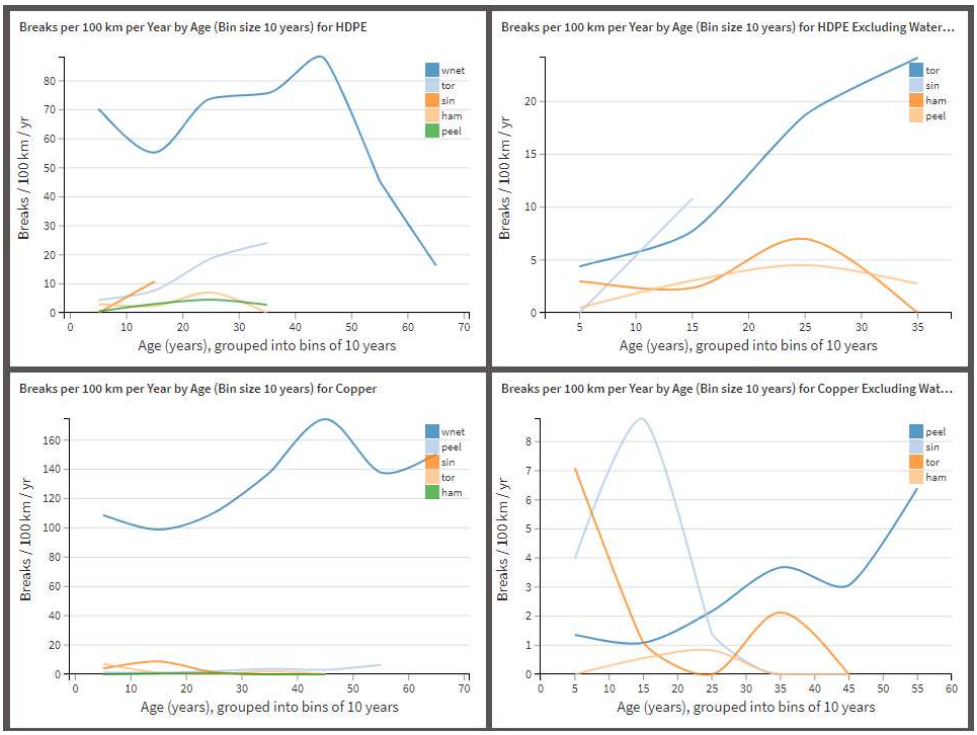


Figure 37: Break rates by age per utility, grouped into bins of 10 years, both with (left) and excluding (right) Waternet, whose break rate for these materials is out of scale with other utilities.

One plausible objection to this trend relates to the relationship between pipe age and the installation dates. Certain utility data contributions span relatively short durations of 10 to 20 years. For such groups, the pipes in certain age bins would have had different manufacturing and installation processes. Take as a hypothetical example a utility which shared data from 2010 to 2019 only. Pipes installed in the year 1970 would appear only in the 60 to 69 years age bin. The 60 to 69 years age bin would only include any entries from pipes installed between 1961 and 1979. The 40 to 49 years age bin would include only pipes installed between 1981 and 1999, which may have followed different manufacturing and/or installation standards. To address this objection, further analysis is provided in Figure 38. This shows data exclusively from long-running data contributions (from Toronto Water and the Region of Peel), where entire pipe cohorts can be tracked through multiple decades of age. The peak-valley trend remains clear: if anything, it is more pronounced within these more homogeneous grouping of pipe.

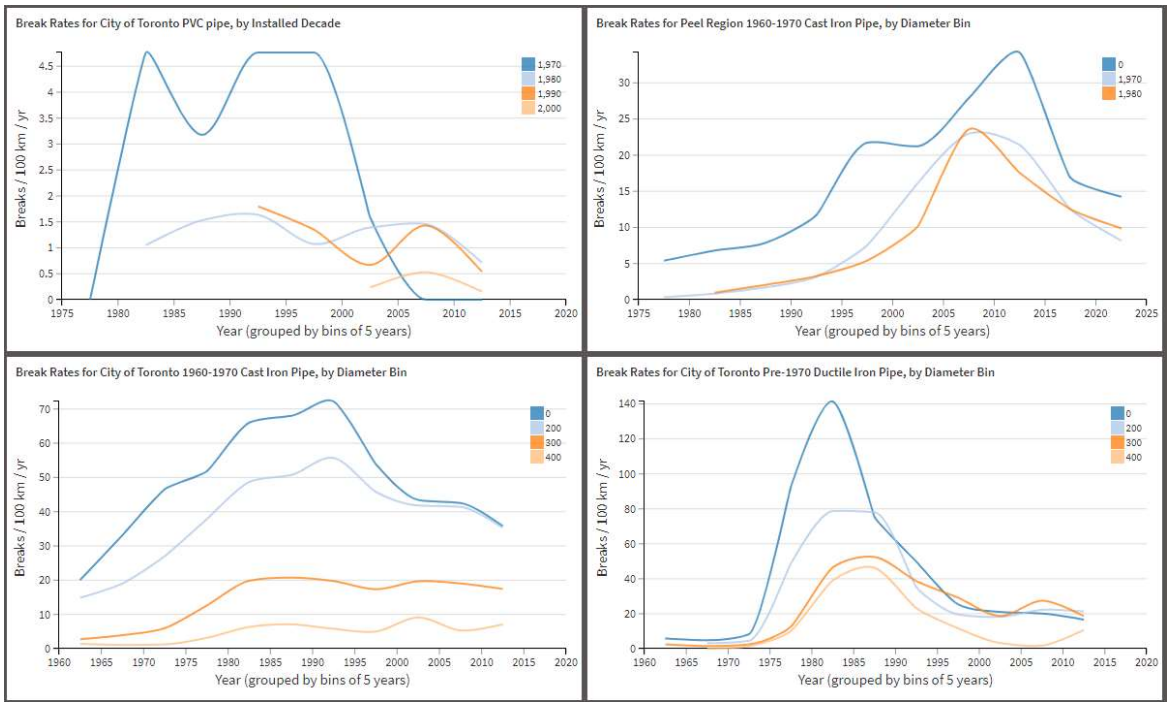


Figure 38: Break rates by year, for various cohorts of pipe in long-running data sets.

These various presentations of the data show a wide range of scales (horizontally and vertically) and clarity of the peak-valley pattern. The pattern is present consistently enough to give confidence that it is not simply an artifact arising out of random variation in a small sample size. While running against the grain of intuition, the peak–valley patterns appear to a degree of consistency that suggests a genuine physical phenomenon is driving it.

4.3 Exploratory Modeling

This section describes the results of the exploratory modeling methodology described in Chapter 3.4.3.

The data set was highly imbalanced, with 0.9% of the Pipe-Year samples exhibiting breaks. This imbalance means that simple accuracy would be a poor metric for model performance. To illustrate this point, a classifier that simply said “pipes never break” would be 99.1% accurate. Models were trained using an inverse class frequency weighting on Break samples.

4.3.1.1 Exploratory Model 1: Random Sample Selection for Train / Test Split

The first set of models used a random selection for the train / test split. While relatively common in the literature reviewed in Chapter 2.3.2, this method allows test samples to be more recent than the training samples, introducing a risk of information leakage from the future to the past. To ensure equal impact of the various utility data sets, the training and test samples were all drawn from the period of 2010 through 2014 when all the data sets overlap. A random sampling of approximately one million samples was used, drawn equally from the six datasets. The chart of performance vs time for all the trained models is shown in Figure 39.

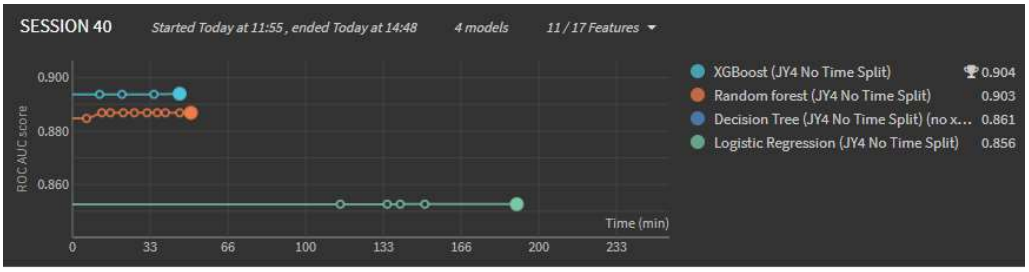


Figure 39: Training performance for baseline models with random train / test assignment.

The top model based on the AUC metric was the Gradient Boosted Tree with Max Depth 5. This agrees with the observation of Snider & McBean (2020b) that boosted decision tree classifiers are well suited to this problem. It achieved an AUC of 0.904, and a Lift at 10% of 6.45. This Lift metric means that by selecting 10% of the population of pipes for replacement using this model, they could capture 64.5% of the next year's breaks.

A comparison of the models' performance across the various dataset subpopulations is provided in Figure 40, with the decision threshold tuned to select 10% of the Segment-Year tuples as having breaks (as per our selected Lift threshold). This shows relatively consistent performance across the six datasets whether measured by AUC or Lift at 10%.

6 modalities for **dataset**, computed on **200k** rows (sample of test set).

Modality	Actually true		Predicted true		Metric: ROC AUC	Lift
100 %	1 %		10 %		0.9043	6.45
17 % — aw	0 %	↓1 %	2 %	↓8 %	0.8448	4.09
17 % — sin	0 %	↓1 %	0 %	↓10 %	0.8700	5.97
17 % — peel	0 %	↓1 %	3 %	↓7 %	0.9408	7.94
17 % — ham	1 %		7 %	↓3 %	0.8814	5.95
17 % — tor	2 %	↑1 %	26 %	↑16 %	0.8370	4.49
17 % — wnet	2 %	↑1 %	23 %	↑13 %	0.8815	5.20

Figure 40: Performance metrics across all utility data sets for gradient boosted tree, the top performing baseline model with random sampling for train and test sets.

4.3.1.2 Exploratory Model 2: Time Dependent Selection of Train / Test Split

The next set of models required that the test set fall strictly after the training set. A similar sampling approach was used for Exploratory Model 1, with the exception that the Train data was drawn exclusively from 2010 to 2012, and the test data was drawn exclusively from 2013 to 2014. This ensures that the Test data falls strictly after the Train data, preventing the possibility of information leakage from the future to the past.

As with the random assignment, the top performing model by AUC was the gradient boosted tree, as shown in Figure 41. The AUC performance of 0.892 was slightly worse than the 0.904 achieved by the same model when the times of the training and test set were intermingled. This was also true of each other model tested (Random Forest, Decision Tree, and Logistic Regression). This consistent effect suggests that there is indeed a small effect of information leakage from the future into the past for pipe failure prediction when the training and test sets are drawn from the same time period. Future research into pipeline failure forecasting would be advised to consider this, and to follow a practice of time-based separation of the training and test sets.

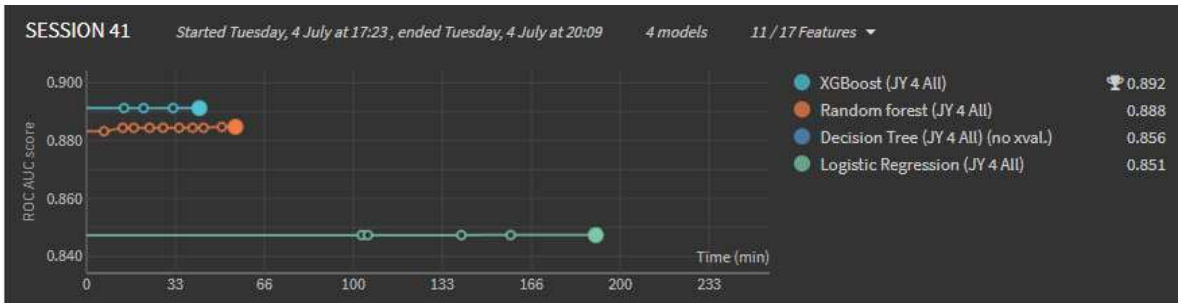


Figure 41: Training performance for baseline models with the test set falling strictly later than the training set.

The top performing model was once again the gradient boosted tree, this time with Max Depth of 4. The AUC and Lift at 10% metrics for this model are shown in Figure 42. As was seen in the AUC scores across the different models, the overall AUC and Lift performance are slightly worse than the top performing model when no time separation was enforced between the train and test sets (shown in Figure 40). Enforcing the time separation degraded performance as measured by AUC on each dataset, and degraded the performance as measured by Lift at 10% for five of the six datasets [peel, sin, ham, tor, wnet]. This offers further validation for the observation of a minor information leakage effect.

6 modalities for **dataset**, computed on **200k** rows (sample of test set).

Modality	Actually true		Predicted true		Metric: ROC AUC	Lift
100 %	1.0 %		10 %		0.89	6.27
17 % — sin	0.2 %	↓0.8 %	0 %	↓10 %	0.85	5.31
17 % — ham	0.8 %	↓0.2 %	5 %	↓5 %	0.86	5.43
17 % — peel	0.4 %	↓0.6 %	3 %	↓7 %	0.90	7.26
17 % — tor	2.4 %	↑1.4 %	23 %	↑13 %	0.83	4.25
17 % — aw	0.4 %	↓0.6 %	3 %	↓7 %	0.81	4.12
17 % — wnet	1.9 %	↑0.9 %	24 %	↑14 %	0.87	4.96

Figure 42: Performance metrics across all utility data sets for gradient boosted tree, the top performing baseline model with the test set falling strictly later than the training set.

The overall performance of this model is encouraging from a practical standpoint. An average Lift at 10% of 6.27 means that if a utility selected 10% of their pipes for replacement or rehabilitation using this algorithm, they could avoid 62.7% of the next year's breaks.

4.3.1.3 Exploratory Model 3: Utility Dependent Selection of Train & Test Sets

Along with extending to future times, a fully generalizable model should also extend to new utilities that did not constitute part of the training data. The third set of baseline models also enforce this requirement in the train / test split of data.

A cross-validation approach was used to achieve this requirement. For each of the six utilities, a training set of roughly 800,000 samples was drawn from the 2010 to 2012 data from the other five, as equally as possible from the utilities. A test set of up to 200,000 samples was drawn from the 2013 and 2014 data for the remaining target utility. The time periods were kept consistent with the other exploratory models to ensure good comparability. Most data sets did not have 200,000 Segment-Year tuples available in the 2013 and 2014 data, in which case a smaller test set was used.

Table 13: Performance metrics for leave-one-utility-out cross validation of baseline models.

Dataset	Test Samples	Top Model	ROC AUC	Lift
Tor	91,406	Random Forest	0.770	3.32
Peel	100,931	Boosted Trees	0.871	5.88
Sin	200,161	Random Forest	0.809	4.06
Wnet	199,650	Boosted Trees	0.792	4.00
Ham	68,730	Random Forest	0.834	4.36
Aw	133,098	Logistic Regression	0.785	3.65
Average	132,329		0.810	4.21

The average performance across the datasets was an AUC of 0.810 and a Lift at 10% of 4.21. This performance is a substantial decline from the AUC of 0.89 and Lift of 6.27 achieved when the train and test data were both drawn from all the utilities. Performance as measured by both AUC and Lift declined for each data set. This suggests that each dataset includes unique information not captured by the others.

The average performance of these models offers a reasonable estimate of the expected performance of Exploratory Model 2 were it to be applied to future data from a new utility. This is a likely means

of application of such an algorithm in practice, particularly among utilities that do not make a practice of tracking pipe failure history in a structured manner.

4.3.1.4 Exploratory Model 4: Train and Test on Individual Utilities

The last set of baseline models is intended to represent a single utility acting on their own, training and testing on their own data alone. This approach is only available to utilities with a relatively large number of segments, as well as a history of tracking pipe failures and allocating them to pipe segments. These conditions are true for all of the utilities participating in this study but are relatively uncommon among the over 30,000 water utilities operating worldwide.

To facilitate comparison with the previous baseline models, the training and test sets are drawn from the same portions of the data. Training data is drawn from the years 2010 through 2012, and test data from 2013 to 2014. For each of the six utility-specific training and test sets, the maximum dataset size was set to $1/6^{th}$ the size used for all of the utilities. This made the number of samples used in these utility-specific tests approximately the same as the number of samples from each utility in the multi-utility tests. For utilities which did not track pipe rehabilitation (aw and peel), this feature was removed.

Table 14: Performance metrics for individual utility baseline models.

Dataset	Test Samples	Top Model	ROC AUC	Lift
Tor	33,306	Boosted Trees	0.831	4.34
Peel	33,044	Boosted Trees	0.909	7.43
Sin	33,342	Random Forest	0.861	6.04
Wnet	33,126	Boosted Trees	0.882	5.33
Ham	33,221	Boosted Trees	0.855	5.09
Aw	33,176	Boosted Trees	0.835	4.32
Average	33,203		0.862	5.43

For all utilities, the individual utility performance was superior to the utility split performance. This result supports the conclusion of Snider & McBean, (2020b) that for utilities with over 10,000 segment-years available, training on that utility's data outperformed a simple rule-based classifier.

This finding suggests that even manual data cleansing is insufficient for full generalization. The result may simply be a matter of too few contributing utilities, or there could be important factors which could differentiate between utilities but are not captured in the data set. Potential factors can be divided into three categories:

- Physical factors which can be represented by measurements from public data sources
    - Temperatures (avg, max, min, range, etc.)
    - Precipitation
    - Typical soil types (sandy, clay, rocky, etc.)
    - Typical surface water acidity (also known as water hardness or PH)
    - Urban vs rural environment
- Physical factors which could be manually obtained from a utility
    - Typical depth of soil cover over pipelines
    - Frequency of proactive leak detection activities
- Information gathering differences among the utilities
    - Whether proactively detected leaks are included in the data set
    - Terminology differences preventing generalization of key fields (material, joints, etc.)
    - Data quality issues that are utility-specific but consistent within that utility

4.3.1.5 Exploratory Model 5: Five-year Future Failure Window Target Variable

An alternative target variable was also tested. Pipeline replacement and rehabilitation programs typically take several years to plan and complete. A five-year time horizon for future breaks was also considered. The methodology was adapted from Exploratory Model 2, with two modifications. First, the target variable was changed to whether a failure will occur on that pipe segment within a five-year window (the current year plus four years into the future). Second, the time window for selecting test and training samples was adjusted for one utility (Toronto Water) to ensure that a full five years of future data was available for calculating the target variable. This data set ends in 2014, so the test set was drawn from the latest two years for which a full five-year window can be calculated (2009 and 2010), and the training set from the three years prior to that (2006 to 2008).

The gradient boosted tree was once again the top performing model, this time with a maximum depth of 5. The dataset was somewhat more balanced than in other models, with 4% of the samples

including a break. The model performance as measured by AUC and Lift at 10% is not substantially different from Exploratory Model 2.

6 modalities for **dataset**, computed on **200k** rows (test set).

Modality	Actually true	Predicted true	Metric: ROC AUC	Lift
100 %	4 %	10.6 %	0.90	6.11
17 % — tor	8 % ↑4 %	25.5 % ↑14.9 %	0.85	4.11
17 % — peel	1 % ↓3 %	2.7 % ↓7.9 %	0.92	7.08
17 % — ham	3 % ↓1 %	8.9 % ↓1.7 %	0.89	5.69
17 % — aw	2 % ↓2 %	2.6 % ↓8 %	0.86	4.80
17 % — sin	1 % ↓3 %	0.5 % ↓10.1 %	0.87	5.73
17 % — wnet	7 % ↑3 %	23.6 % ↑13 %	0.89	4.80

Figure 43: Performance metrics for five-year future failure prediction across all utility data sets for gradient boosted tree, the top performing baseline model with the test set falling strictly later than the training set.

4.3.1.6 Summary of Exploratory Segment-Year Model Performance

Baseline performance was measured for the six datasets using several different rules for selecting the training and test sets. These results, presented in Table 15, each represent different potential usage scenarios for the models.

Table 15: Summary of exploratory modeling results

	1: Fully Random		2: Time Split		3: Utility Split		4: Individual Utility		5: Break in 5 Yrs	
Dataset	AUC	Lift	AUC	Lift	AUC	Lift	AUC	Lift	AUC	Lift
Tor	0.84	4.49	0.83	4.25	0.77	3.32	0.83	4.34	0.85	4.11
Peel	0.94	7.94	0.90	7.26	0.87	5.88	0.91	7.43	0.92	7.08
Sin	0.87	5.97	0.85	5.31	0.81	4.06	0.86	6.04	0.87	5.73
Wnet	0.88	5.20	0.87	4.96	0.79	4.00	0.88	5.33	0.89	4.80
Ham	0.88	5.95	0.86	5.43	0.83	4.36	0.86	5.09	0.89	5.69
Aw	0.85	4.09	0.81	4.12	0.79	3.65	0.84	4.32	0.86	4.80
Average	0.876	5.61	0.85	5.22	0.810	4.21	0.862	5.43	0.88	5.37

The best average performance was achieved by the classifier that permits the train and test set time periods to overlap. Among the remaining train/test split rule sets which prevent this, the top performance was achieved by training separate classifiers on each utility individually. The worst performance achieved was for training a classifier on data from all but one utility, and then testing it on data from that utility.

The Random Assignment baseline introduces information leakage from the future into the past. It is provided only for the purpose of comparison to other studies that may overlook this issue.

The two baselines which most realistically simulate real world use are the Individual Utility and Utility Split rules, both with time separation of the train / test sets. The Individual Utility rule represents an individual utility with their own large dataset training a model on their own data and then applying it to their own future predictions. This will be the basis for the Isolated data inclusion scheme used in testing the final model in this study. The Utility Split rule represents a pre-trained model being prepared, and then applied at a later data to a new utility. This will be the basis for the Leave One Group Out (LOGO) data inclusion scheme used in testing the final model in this study.

4.4 Chapter Summary

This chapter described the results of the exploratory analysis, which was conducted following the methodology described in Chapter 3.4. Basic descriptive statistics of the data were provided, as well as a confirmation of the appropriateness of normalizing failure rates by pipe length. The impact on failure rates was examined for individual features expected to be predictive based on the literature review, such as diameter, material, and age. Two findings of interest emerged, likely due to the large size and diversity of the dataset: (1) a strong inverse relationship between diameter and failure rate, and (2) a pattern of peaks and subsequent declines in the failure rate as age increased. The results of exploratory modeling were also described, which broadly agreed with past results from the literature review. The exploratory modeling showed that models perform worse when tested on data from future time periods or other utilities, validating the direction of research towards a generalized model which can extrapolate forward in time and to new utilities.

Chapter 5 describes the results obtained from the generalized machine learning model for pipe failure prediction, following the methodology described in Chapter 3.5. The results of each of the

three layers of the model are described individually, to provide confirmation that each layer is serving its intended function effectively. A detailed analysis of the performance of the full model is then provided, including performance on various subpopulations and the impact of particular predictive features.

Chapter 5
Results: Generalized Machine Learning Model

This chapter presents the performance of the final generalized machine learning model.

5.1 Layer 1: Feature Preprocessing

This section provides the results obtained by the feature preprocessing layer. It shows the intermediate results fed forward into the failure classification model layer. The purpose of showing these results is to evaluate the effective performance of this layer.

5.1.1 Infer Units for Objective Numerical Features

As described in Chapter 3.5.3, the feature preprocessing layer aims to prepare the feature vector x by inferring the units of numerical features and encoding categorical features.

The units inference was performed by testing all plausible combinations of units among the six utilities contributing data. Two numerical fields were converted: length and diameter. The aggregate metrics before conversion were as follows:

```
Diameter summary:
                mean   median          std
dataset
aw          9.462854      8.0     6.795564
ham       226.384828    150.0   160.504489
peel      253.592924    200.0   184.815599
sin       228.570478    150.0   211.028601
tor       237.364365    150.0   199.035686
wnet      419.623208    150.0   490.320509

Length summary:
                mean       median          std
dataset
aw        239.717502   159.251702   279.848256
ham        60.934126    20.070000    90.272359
peel      106.095258    65.240000   155.239076
sin        40.465063    10.193719    94.657403
tor       134.513141    96.600000   148.632137
wnet       24.352321     6.040000    47.979690
```

The top 10 unit assignment sets, and the resulting sum of scale variance, were as follows:

Diameter Units	Length Units	Sum Variance
(in, mm, mm, mm, mm, mm)	(ft, m, m, m, m, m)	1.458187
(in, mm, in, mm, in, mm)	(ft, m, ft, m, ft, m)	1.571404
(in, mm, in, in, in, mm)	(ft, m, ft, ft, ft, m)	2.243774
(in, mm, in, mm, in, in)	(ft, m, ft, m, ft, ft)	2.293011
(in, in, in, mm, in, mm)	(ft, ft, ft, m, ft, m)	2.306178
(in, mm, mm, mm, in, mm)	(ft, m, m, m, ft, m)	2.570075
(in, mm, mm, mm, in, in)	(ft, m, m, m, ft, ft)	2.658917
(in, mm, mm, in, in, mm)	(ft, m, m, ft, ft, m)	2.727322
(in, in, mm, mm, in, mm)	(ft, ft, m, m, ft, m)	2.882029

Based on a manual review of the data, the top performing set of unit assignments was the correct set of assignments. It is worth noting that the number of datasets in use for this unit inference model was relatively small. Exploring whether this approach works on larger collections of datasets and in other domains is left as an area for future research.

5.1.2 Encode Categorical Features

The specific encodings applied in this study were averages of the following numerical features, for the [utility-categorical feature] tuples shown. The encodings were developed using the averages from the training data only (i.e., the test data was excluded) to prevent information leakage from the future into the past. Note that "Had Break This Year" is a binary (0 or 1) feature; hence, the average represents the proportion of samples which experienced a break.

- Utility-Material: Diameter, Length, Install Year, Had Break This Year
- Utility-Lining: Diameter, Length, Install Year, Had Break This Year
- Utility-Corrosion_Protection: Diameter, Length, Install Year, Had Break This Year
- Utility-Joints: Diameter, Length, Install Year, Had Break This Year
- Utility-Main_Type: Diameter, Length, Install Year, Had Break This Year

These encodings of categorical features result in a total of 31 features being used by the machine learning model in Layer 2, as compared to 141 features when dummy encoding was applied.

The effectiveness of the feature encodings can be measured by their impact on the performance of the machine learning classifiers trained in Layer 2.

5.2 Layer 2: Failure Classification

Note that the performance in this section cannot be compared exactly to the Exploratory Modeling results in Chapter 4.3, as a different Train and Test sample set were used. Specifically, this set took a random sampling from a longer duration in time for the training set. This is expected to result in slightly worse performance on the test set, but better generalizability to test sets not from time periods adjacent to the training set.

Performance was measured by four metrics, including the two primary metrics for this study (ROC AUC and Lift at 10%) and two secondary metrics (Log Loss, and Maximum F1 Score). Tests were done for each of the three data inclusion schemes described in Chapter 3.5.4, both on the original data and the data after preprocessing by Layer 1.

5.2.1 Calculation of Relative Performance Between Models

Also presented are the relative performance of LOGO versus both the Isolated and Inclusive schemes. Relative performance was measured as follows:

- **ROC AUC:** (LOGO – 0.5) / (Comparable – 0.5). Since a random estimator would achieve an AUC score of 0.5, this sets the practical floor for performance. Subtracting 0.5 from each score provides an estimate of the relative performance gain or loss.
- **Lift at 10%:** LOGO / Comparable. Simple proportionate score.
- **Log Loss:** Comparable / LOGO. Log Loss is negatively oriented, with 0 being a perfect score.
- **F1 Score:** LOGO / Comparable. Simple proportionate score.

Wherever a "Gain" is referred to as a relative performance measure, this is calculated as (Score – Comparable) / Comparable.

5.2.2 Results on Raw Data

Results of the three data inclusion schemes prior to application of the preprocessing layer are presented in Table 16 below. The results were unsurprising, with the Inclusive scheme offering the best performance by all metrics. The worst performance across all metrics was achieved by the LOGO scheme, with Isolated falling in the middle. This top performance of the inclusive scheme

confirms that there is in fact a degree of generalizability in the data, since access to the break history from other utilities does, on average, improve performance compared to using each utility's data in isolation.

Table 16: Failure classification model results on raw data.

	Results on Raw Data			LOGO Performance Vs	
	Isolated	Inclusive	LOGO	Isolated	Inclusive
By ROC AUC	0.870	0.885	0.782	76.2%	73.2%
By Lift at 10%	4.989	5.650	4.232	84.8%	74.9%
By Log Loss	0.424	0.399	0.519	81.7%	76.9%
By F1 Score	0.269	0.317	0.241	89.6%	76.0%
Average Relative Score Across Metrics				83.1%	75.3%

The average relative performance of the LOGO scheme across these four metrics was 83.1% as strong as the Isolated scheme, and 75.3% as strong as the top performing Inclusive scheme. This result confirms that differences in the data from contributing utilities are preventing full generalization.

5.2.3 Results on Preprocessed Data

Results of the three data inclusion schemes after application of the preprocessing layer are presented in Table 17 below. Once again, the Inclusive scheme is the strongest performing by all metrics. The performance of the LOGO and Isolates schemes are now mixed, with each scoring last in two out of four metrics.

Table 17: Failure classification model results on preprocessed data

	Results on Preprocessed Data			LOGO Performance Vs	
	Isolated	Inclusive	LOGO	Isolated	Inclusive
By ROC AUC	0.871	0.889	0.863	97.8%	93.3%
By Lift at 10%	5.026	5.784	5.092	101.3%	88.0%
By Log Loss	0.420	0.399	0.377	111.4%	105.8%
By F1 Score	0.267	0.322	0.271	101.5%	84.2%
Average Relative Score Across Metrics				103.0%	92.8%

The performance gap between the different LOGO and the other data inclusion schemes has shrunk considerably. The average relative performance of the LOGO scheme across the four metrics is now 103% as strong as the Isolated scheme, and 92.8% as strong as the top performing Inclusive scheme.

5.2.4 Relative Performance Before and After Preprocessing

Comparative results for each data inclusion scheme are presented in Table 18 below. The preprocessing has had a negligible impact on the performance of the Isolated and Inclusive schemes. The performance of the LOGO scheme, however, improved by between 12% and 29% on each of the performance metrics used.

Table 18: Failure classification model performance gains introduced by preprocessing layer.

	Isolated			Inclusive			LOGO		
	Raw	Preproc.	Gain	Raw	Preproc.	Gain	Raw	Preproc.	Gain
By ROC AUC	0.870	0.871	0.3%	0.885	0.889	1.0%	0.782	0.863	28.7%
By Lift at 10%	4.989	5.026	0.7%	5.650	5.784	2.4%	4.232	5.092	20.3%
By Log Loss	0.424	0.420	0.9%	0.399	0.399	0.0%	0.519	0.377	27.4%
By F1 Score	0.269	0.267	-0.7%	0.317	0.322	1.6%	0.241	0.271	12.4%

Table 19 presents a direct comparison of the LOGO performance on the raw and preprocessed data. It is shown as a relative performance to both the Isolated and the Inclusive schemes. Measurement of relative performance is as described in Chapter 5.2.1. Averaged across the performance metrics, the preprocessing layer closes the gap versus the Isolated scheme by nearly 20% (to the point of slightly outperforming) and closes the gap by over 17% on the Inclusive scheme.

Table 19: Relative performance gains introduced by preprocessing layer in the LOGO data inclusion scheme, as compared to Isolated and Inclusive schemes.

	LOGO vs Isolated			LOGO vs Inclusive		
	Raw	Preproc	Gain	Raw	Preproc	Gain
By ROC AUC	76.2%	97.8%	21.6%	73.2%	93.3%	20.1%
By Lift at 10%	84.8%	101.3%	16.5%	74.9%	88.0%	13.1%
By Log Loss	81.7%	111.4%	29.7%	76.9%	105.8%	29.0%
By F1 Score	89.6%	101.5%	11.9%	76.0%	84.2%	8.1%
Average Across Metrics	83.1%	103.0%	19.9%	75.3%	92.8%	17.6%

The final comparison for Layer 2, as shown in Table 20, is between the LOGO scheme on the Preprocessed data with the Isolated and Inclusive schemes on the original raw data. These metrics compare the relative performance that a utility could expect were they to use the pretrained generalized model, versus participating in a multi-utility study or training their own model on their own data.

Table 20: Final LOGO model performance compared with Isolated and Inclusive model performance on raw data.

	LOGO on	Vs. Isolated on Raw		Vs. Inclusive on Raw	
	Preprocessed	Score	Relative	Score	Relative
By ROC AUC	0.863	0.870	98.1%	0.885	94.3%
By Lift at 10%	5.092	4.989	102.1%	5.650	90.1%
By Log Loss	0.377	0.424	112.5%	0.399	105.8%
By F1 Score	0.271	0.269	100.7%	0.317	85.5%
Average Across Metrics			103.3%		93.9%

5.3 Layer 3: Calibration

This section provides the results obtained by the final calibration layer. The outcomes of each of the two calibration steps are shown. First is the creation of a calibrated probability of failure estimate using isotonic regression. Second is the estimation of the expected number of failures in a segment by application of a linear scaling to account for multiple failures in the same pipe within a five-year period. The expected number of failures are then summed up over a cohort of segments for the purpose of performance evaluation.

5.3.1 Impact of Isotonic Regression

The effectiveness of the Calibration layer is measured in two ways. The impact of applying Isotonic Regression is measured on the Calibration layer on the metrics previously discussed, as well as on the Calibration Loss metric described in Chapter 3.5.5. These results are presented in Table 21.

Table 21: Performance of models before and after application of isotonic regression to calibrate probability estimates.

	Isolated		Inclusive		LOGO	
	Before	After	Before	After	Before	After

By ROC AUC	0.871	0.872	0.889	0.890	0.863	0.864
By Lift at 10%	5.026	5.084	5.784	5.801	5.092	5.092
By Log Loss	0.420	0.115	0.399	0.110	0.377	0.117
By F1 Score	0.267	0.269	0.322	0.325	0.271	0.272
By Calibration Loss	0.244	0.000	0.229	0.000	0.201	0.000

Three of the metrics (AUC, Lift at 10%, and F1 Score) are largely unaffected by the application of isotonic regression. This is a consequence of the fact that these metrics rely exclusively on the relative probability estimates between samples. Since the ordinality of the probability scores is preserved by the calibration, these scores are unchanged. Minor variations are explained by the presence of tied probability scores among samples, which are relatively common in decision-tree based models. These tied scores may be sorted in a random manner, leading to slight variations in the scoring before and after application of isotonic regression. The Calibration Loss reduces to zero for all data inclusion schemes, which is an expected consequence of applying isotonic regression on the same binning scheme used to calculate the calibration loss. The Log Loss metric improves significantly, as this metric is impacted by both the relative rankings and the accuracy of the probability estimate itself.

5.3.2 Ability to Predict Total Breaks in a Cohort

The last evaluation of the calibration layer is the ability of the model to predict the actual number of breaks in a future period on a cohort. Performance was measured by comparing the Expected Number of Failures (ENoF) summation to the actual total number of breaks in cohort over the ensuing five-year period, as shown in Figure 44.

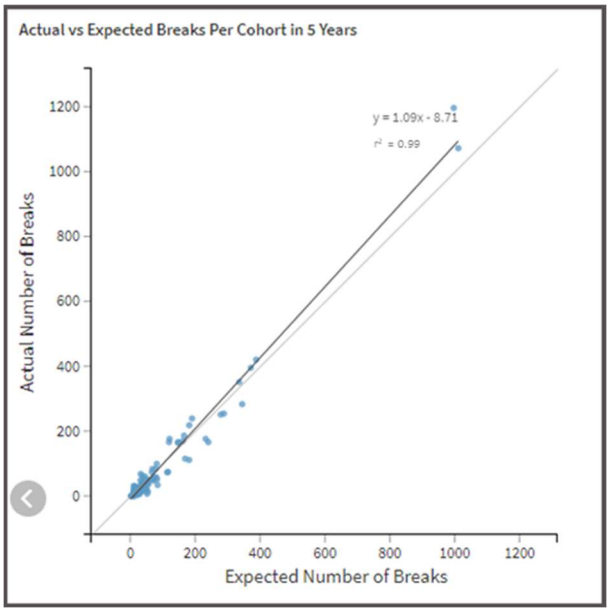


Figure 44: Expected Number of Failures from model, aggregated to cohorts, compared to actual failures in those cohorts.

As a reference point, the performance of a simple "Predict the same number of breaks as occurred in the previous five-year period" model (referred to as the Prior Period model) was also measured. Comparative charts are presented in Figure 45, and Table 22 shows the results using the metrics described in Chapter 3.6.3.

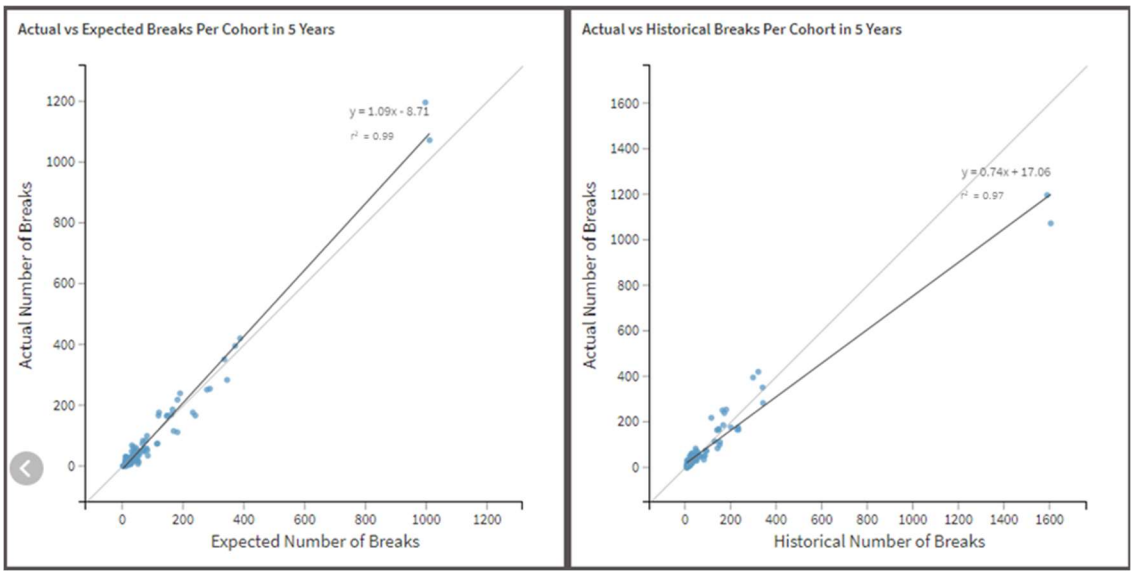


Figure 45: Comparative plots of Expected Number of Failures from the model and Previous Period Failures each plotted versus actual failures in next period.

Table 22: Performance metrics of Expected Number of Failures from the model and Prior Period Failures used to forecast number of failures in the upcoming period.

	MAE	Bias	MSE	RMSE	Linear Regression	R^2
ENoF	18.11	1.30	899	29.98	y = 1.09x - 8.71	0.99
Prior Period	26.19	4.65	4669	68.33	y = 0.74x + 17.06	0.97

The ENoF summation outperformed the Prior Period model by all metrics. It is worth noting that the Prior Period model is limited in practical application, since it can only be used for the upcoming five-year period. The ENoF from the model, however, can be projected forward simply by incrementing the Age features of each entry in the test set by the desired number of years, and then re-running Layers 2 and 3.

5.4 Detailed Model Performance Analysis

This section presents performance details for the full proposed model. This consists of subpopulation analysis to evaluate performance of the model within various subgroups of the data, as well as individual feature explanatory analysis.

5.4.1 Subpopulation Analysis

This subsection presents analysis of model performance on various subpopulation groups among the test set. In each instance, the decision threshold for the model was tuned to pick the 10% highest risk pipes from the overall population for replacement.

5.4.1.1 By Contributing Utility

A summary of performance on the data from the six utilities is presented in Table 23, with the model decision threshold tuned to select approximately 10% of all pipes (across all utilities) for replacement.

Table 23: Subpopulation analysis of final model performance by contributing utility, using a common decision threshold to select 10% of pipes across the full data set.

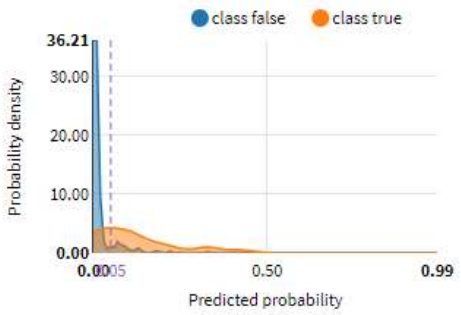
Contributing Utility	ROC AUC	Lift	Accuracy	Precision	Recall
peel	0.879	6.51	0.95	0.13	0.45
wnet	0.880	4.84	0.93	0.32	0.34
ham	0.859	5.50	0.94	0.25	0.36
sin	0.796	3.95	0.99	0.12	0.07
tor	0.796	3.31	0.65	0.16	0.81
aw	0.812	4.27	0.95	0.11	0.25

An alternate analysis is provided in Table 24 with the decision thresholds tuned separately for each utility, to select approximately 10% of the pipes within that utility. Note that the ROC AUC does not change, as this is independent of the decision threshold. The other metrics do change, including Lift. While both Lift figures represent "Lift at 10%," the definition has changed from "at 10% of all pipes" to "at 10% of the pipes within this utility."

Table 24: Subpopulation analysis of final model performance by contributing utility, using varied decision thresholds to select 10% of pipes for each utility.

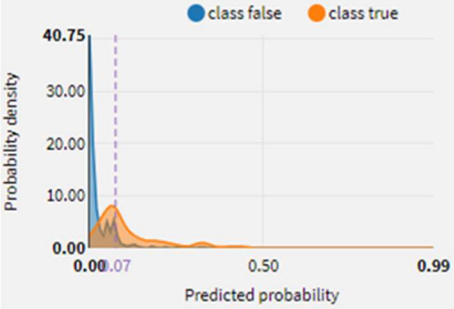
Contributing Utility	ROC AUC	Lift	Accuracy	Precision	Recall
peel	0.879	6.23	0.89	0.09	0.69
wnet	0.880	5.13	0.91	0.28	0.44
ham	0.859	5.82	0.91	0.19	0.53
sin	0.796	4.14	0.92	0.04	0.34
tor	0.796	3.25	0.87	0.26	0.34
aw	0.812	4.69	0.91	0.08	0.39

Further details for each contributing utility are provided in Figure 46 through Figure 51, also using the decision thresholds tuned to select approximately 10% of pipes per utility. These figures provide both a confusion matrix and a density plot, which shows the probability density of the true (i.e., break) and false (i.e., no break) populations by predicted probability. Greater separation between the two groups indicates better model performance on that subpopulation.



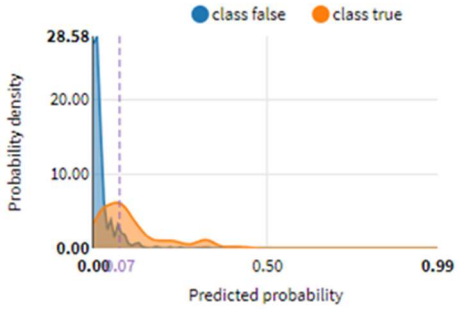
	Predicted `true`	Predicted `false`	Total
Actually `true`	332	147	479
Actually `false`	3384	29544	32928
Total	3716	29691	33407

Figure 46: Subpopulation analysis for utility contributor Peel.



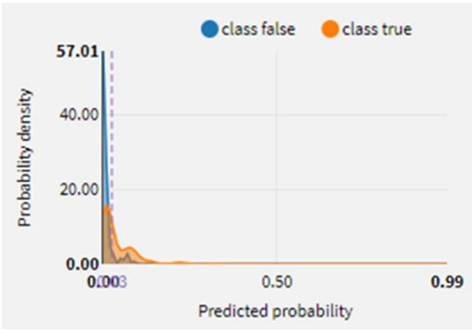
	Predicted `true`	Predicted `false`	Total
Actually `true`	784	1012	1796
Actually `false`	2054	29509	31563
Total	2838	30521	33359

Figure 47: Subpopulation analysis for utility contributor Waternet.



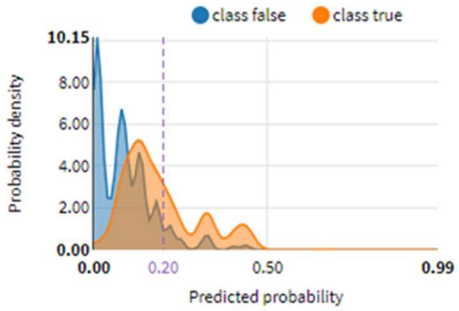
	Predicted `true`	Predicted `false`	Total
Actually `true`	582	521	1103
Actually `false`	2429	29675	32104
Total	3011	30196	33207

Figure 48: Subpopulation analysis for utility contributor Hamilton.



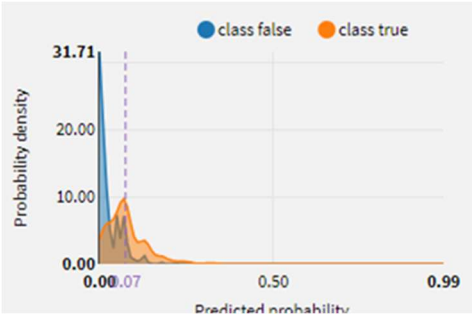
	Predicted true	Predicted false	Total
Actually true	113	221	334
Actually false	2597	30260	32857
Total	2710	30481	33191

Figure 49: Subpopulation analysis for utility contributor Singapore.



	Predicted true	Predicted false	Total
Actually true	902	1735	2637
Actually false	2589	27948	30537
Total	3491	29683	33174

Figure 50: Subpopulation analysis for utility contributor Toronto.



	Predicted true	Predicted false	Total
Actually true	228	351	579
Actually false	2544	29913	32457
Total	2772	30264	33036

Figure 51: Subpopulation analysis for utility contributor American Water.

The lowest performing utilities were Sin and Tor, but for opposite reasons. The Sin network has both the lowest rate of breaks and the lowest prevalence of break risk indicators among the features. The decision threshold must be tuned to an extremely low level (2.5% predicted probability of failure) to select an appropriate number of pipes; however, this tuning results in selecting many pipes that will not fail. The Tor network is at the opposite end of the spectrum, including both the highest rate of breaks and the greatest prevalence of risk indicators. The decision threshold must be tuned

substantially higher (20% predicted probability of a failure) to select an appropriate number of pipes; however, this still omits many high risk pipes.

These two utilities represent the "outliers" on the spectrum of break prevalence. The remaining four fall in between, with decision thresholds between 5% and 7.5% resulting in appropriate numbers of pipes. It is not surprising that the "Leave One Group Out" method performs worst when applied to utilities on the edges of the performance spectrum, as it is well known that machine learning models interpolate better than they extrapolate.

5.4.1.2 By Material

For the purposes of subpopulation analysis, the manual data cleansing on material type was used, which allows for grouping of samples based on their actual material. Note that this manual data cleansing was not used in the model tuning or sample evaluation, but only for the purpose of grouping samples together for subpopulation analysis. As with utility contributors, these are presented both with decision thresholds tuned to select 10% of overall pipes (Table 25), and also to select approximately 10% of the pipes of in that material class (Table 26).

Table 25: Subpopulation analysis of final model performance by pipe material, using a common decision threshold to select 10% of pipes across the full data set.

Material Group	ROC AUC	Lift	Accuracy	Precision	Recall
Ductile Iron	0.84	5.38	0.96	0.16	0.34
Cast Iron	0.79	3.41	0.76	0.18	0.63
PVC	0.91	7.61	0.98	0.32	0.25
Concrete	0.87	5.86	0.99	0.02	0.02
Unknown	0.89	6.81	0.92	0.10	0.65
Steel	0.81	3.06	0.99	0.00	0.00
Copper	0.84	4.36	0.94	0.28	0.22
PE	0.84	3.89	0.92	0.31	0.26
Asbestos Cement	0.69	2.55	0.89	0.14	0.25

Table 26: Subpopulation analysis of final model performance by material, using varied decision thresholds to select approximately 10% of pipes for material.

Material Group	ROC AUC	Lift	Accuracy	Precision	Recall
Ductile Iron	0.84	4.86	0.89	0.09	0.56

Cast Iron	0.79	3.93	0.90	0.28	0.26
PVC	0.91	6.39	0.89	0.10	0.79
Concrete	0.87	7.16	0.95	0.02	0.32
Unknown	0.89	6.51	0.90	0.08	0.72
Steel	0.81	4.43	0.95	0.03	0.19
Copper	0.84	4.81	0.92	0.22	0.29
PE	0.84	4.30	0.91	0.25	0.32
Asbestos Cement	0.69	3.91	0.89	0.14	0.25

The model performance was strongest as measured by both Lift and AUC on PVC, Concrete, and (somewhat surprisingly) where the material was unknown. Weaker performance was observed on asbestos cement and on cast iron. It is plausible that these materials, being both of high average age and having high break rates, would benefit most from allowing inclusion of longer break history records as a feature.

5.4.1.3 By Diameter

Performance by diameter is summarized in Table 27. Diameters were grouped automatically into seven bins of approximately equal population. Square brackets indicate that the bin is inclusive of the value, and round brackets indicate that the bin excludes the value.

Table 27: Subpopulation analysis of final model performance by diameter.

Diameter Bin	ROC AUC	Lift	Accuracy	Precision	Recall
[1, 110]	0.85	4.07	0.93	0.26	0.27
(110, 150]	0.85	4.30	0.82	0.18	0.67
(150, 160]	0.78	3.30	0.91	0.12	0.29
(160, 200]	0.89	5.92	0.94	0.19	0.45
(200, 300]	0.86	5.41	0.93	0.15	0.43
(300, 406]	0.83	5.36	0.96	0.11	0.29
(406, 3300]	0.85	4.54	0.99	0.06	0.06

No noteworthy performance differences are notable between the smaller and larger diameter pipes. A minor anomaly is the performance of pipes between 151 and 160mm in diameter. This small bin essential includes pipe with a nominal diameter of 6 inches. Only one of the contributing utilities uses pipe of this precise size; hence, the Leave One Group Out modeling approach would not have

had any pipe of identical diameter in the training set. This suggests that the model performance may improve if trained using data from at least one utility using pipe of identical diameters.

5.4.1.4 By Length

Performance of the model by length bin is presented in Table 28. Bins were formed automatically to include 10% of the pipe population in each. Square brackets indicate that the bin is inclusive of the value, and round brackets indicate that the bin excludes the value. Only metrics that are independent of the decision threshold are presented, as the class balance is highly dependent on length, making threshold-based metrics such as precision and recall less useful.

Table 28: Subpopulation analysis of final model performance by length.

Length Bin	ROC AUC	Lift
[0.012, 2.27]	0.698	3.43
(2.27, 5.05]	0.775	4.10
(5.05, 8]	0.847	5.65
(8, 14.72]	0.790	3.96
(14.72, 28.10]	0.798	4.22
28.10, 53.40]	0.790	3.35
(53.40, 80.47]	0.772	3.19
(80.47, 113]	0.752	2.94
(113, 187]	0.780	3.30
(187, 5478]	0.796	3.39

Note that the AUC and Lift scores for a given bin show the model's ability to distinguish between the low- and high-risk pipes within that same bin. The lower scores in all bins indicate that the model is less effective distinguishing break rates among pipe of similar length. This finding is not surprising, as length is a major contributor to the model. The average AUC of 0.780 and Lift of 3.75 within these bins offers a reasonable estimate of the model's ability to prioritize pipes of similar length.

5.4.1.5 By Age and Installed Date

Model performance by age is presented in Table 29. The test data was automatically grouped into 10 bins of approximately equal size. Square brackets indicate that the bin is inclusive of the value, and round brackets indicate that the bin excludes the value.

Table 29: Subpopulation analysis of final model performance by pipe age.

Age bin	ROC AUC	Lift	Accuracy	Precision	Recall
[0, 7]	0.83	5.70	0.99	0.20	0.12
(7, 12]	0.9	7.00	0.99	0.32	0.18
(12, 18]	0.88	6.90	0.98	0.34	0.23
(18, 25]	0.85	5.54	0.97	0.21	0.21
(25, 31]	0.82	4.90	0.95	0.18	0.26
(31, 41]	0.83	4.60	0.89	0.18	0.50
(41, 48]	0.86	4.82	0.87	0.20	0.57
(48, 58]	0.82	3.47	0.75	0.22	0.71
(58, 86]	0.77	3.23	0.80	0.16	0.53
(86, 255]	0.77	3.30	0.82	0.10	0.49

Similar performance measures grouped by installed date are presented in Table 30.

Table 30: Subpopulation analysis of final model performance by year of installation.

Installed Date bin	ROC AUC	Lift	Accuracy	Precision	Recall
[1758, 1925]	0.77	3.24	0.80	0.10	0.53
(1925, 1954]	0.8	3.36	0.79	0.20	0.60
(1954, 1964]	0.82	3.53	0.76	0.21	0.69
(1964, 1972]	0.85	4.77	0.87	0.17	0.55
(1972, 1981]	0.82	4.48	0.90	0.16	0.43
(1981, 1989]	0.83	5.30	0.96	0.24	0.25
(1989, 1995]	0.86	5.78	0.97	0.23	0.18
(1999, 2001]	0.89	6.90	0.99	0.34	0.21
(2001, 2006]	0.89	6.72	0.99	0.33	0.18
(2006, 2014]	0.83	5.65	0.99	0.19	0.12

The model is reasonably effective distinguishing between high- and low-risk pipes of similar age or installation vintage. Model performance is somewhat weaker among very new and very old pipes. For very new pipes, this result may be a consequence of a higher proportion of wholly unpredictable failures (installation errors, third party damage, etc.). For very old pipes, this may be an area where the model would have benefited from being permitted to use longer term break history. The model's average Lift of 4.95 across these bins confirms that machine learning models can provide utilities substantial value over age-based replacement programs.

5.4.1.6 By Pipe History

Table 31 presents the performance based on past rehabilitation status.

Table 31: Subpopulation analysis of final model performance by past rehabilitation status.

Past Rehab Status	ROC AUC	Lift	Accuracy	Precision	Recall
FALSE	0.87	5.25	0.93	0.17	0.43
TRUE	0.73	2.47	0.62	0.18	0.74

Performance of the model is markedly worse on pipes with a history of rehabilitation. It is possible that insufficient information regarding the past rehabilitation status is being provided to the model. The type of rehabilitation or the time since rehabilitation may be important information. The long-term impact of rehabilitation on break rates offers an interesting area of potential for future research.

Model performance by break history is presented in two divisions: whether a break occurred in the past five years (one of the features), and whether or not a break occurred at any point in the pipe's records (not a model feature) in Table 32 and Table 33 respectively.

Table 32: Subpopulation analysis of final model performance by whether a break was recorded in the prior five-year period.

Break Past 5 Years	ROC AUC	Lift	Accuracy	Precision	Recall
FALSE	0.84	4.35	0.94	0.14	0.25
TRUE	0.61	1.53	0.45	0.32	0.84

Table 33: Subpopulation analysis of final model performance by whether a break was recorded at any point in the available records.

Any Break in Records	ROC AUC	Lift	Accuracy	Precision	Recall
FALSE	0.81	3.85	0.96	0.07	0.12
TRUE	0.64	1.81	0.56	0.27	0.66

Distinguishing between pipes with similar break history is clearly challenging for the model. Good performance is achieved among pipes with no failures (either in the past five years, or anywhere in their historical records); however, the model struggles to distinguish between high and low risk pipes

that have had a previous or recent failure. This confirms the findings in Snider & McBean (2021) that pipes with previous failures behave markedly differently from pipes with no previous failures.

This presents another interesting area for future research. Models developed specifically for the purpose of predicting future failures among pipes with a previous failure history could be both interesting and of great practical value.

5.4.2 Individual Feature Analysis

5.4.2.1 Feature Importance by Shapley Values

The relative contributions of the 20 most important features, as measured by Shapley values, are shown in Figure 52: Relative contributions of top 20 features in the model as measured by Shapley values.. The contributions are concentrated among a relatively small number of features. In agreement with previous findings in the literature, the significant features include length and diameter (captured jointly in the length to diameter ratio and separately in length), material (as encoded by the historical break rate and installation year), and age. While the dataset (identity of the contributing utility) is not a feature in the model, it is used to direct each sample to the correct model; hence, its relative contribution can also be measured with the Shapley value and was found to be significant. Lesser contributions come from the year of installation and the recent break history.

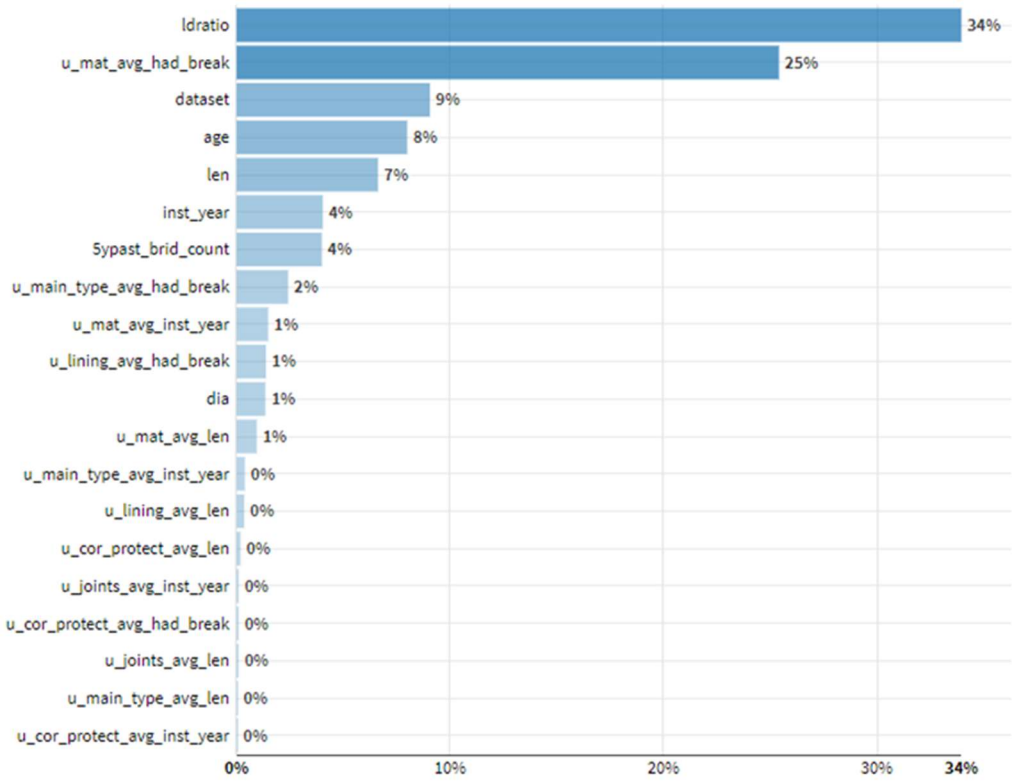


Figure 52: Relative contributions of top 20 features in the model as measured by Shapley values.

Shapley values are computed per sample, and are directional, with positive numbers indicating the value was pushed towards the positive class and negative values indicating the value was pushed towards the negative class. The directional contributions of the top 20 features on a random selection of samples is shown in Figure 53. One aspect of this chart which particularly stands out is the contribution of the break history. While this value is 0 for almost all samples, the cases with larger values (1 or greater) have strongly positive Shapley contributions. Taken together with the overall Shapley contributions in Figure 53, this finding indicates that previous breaks are an infrequent but potent predictor of future breaks.

Figure 53: Directional contributions of top 20 features, as measured by Shapley values, for a random selection of samples from the test set.

A selection of plots of Shapley value contributions are shown for features of interest in Figure 54: Plots of the contribution by features of greatest impact, as measured by Shapley values on a random selection of samples from the test set. For the most part the trends are unsurprising and conform to the expectations from past studies and the exploratory analysis. Confirming the exploratory findings on age discussed in Chapter 4.2.4, the plot of age shows minor overall contributions, with positive contributions peaking around age 40 to 50 and then declining.

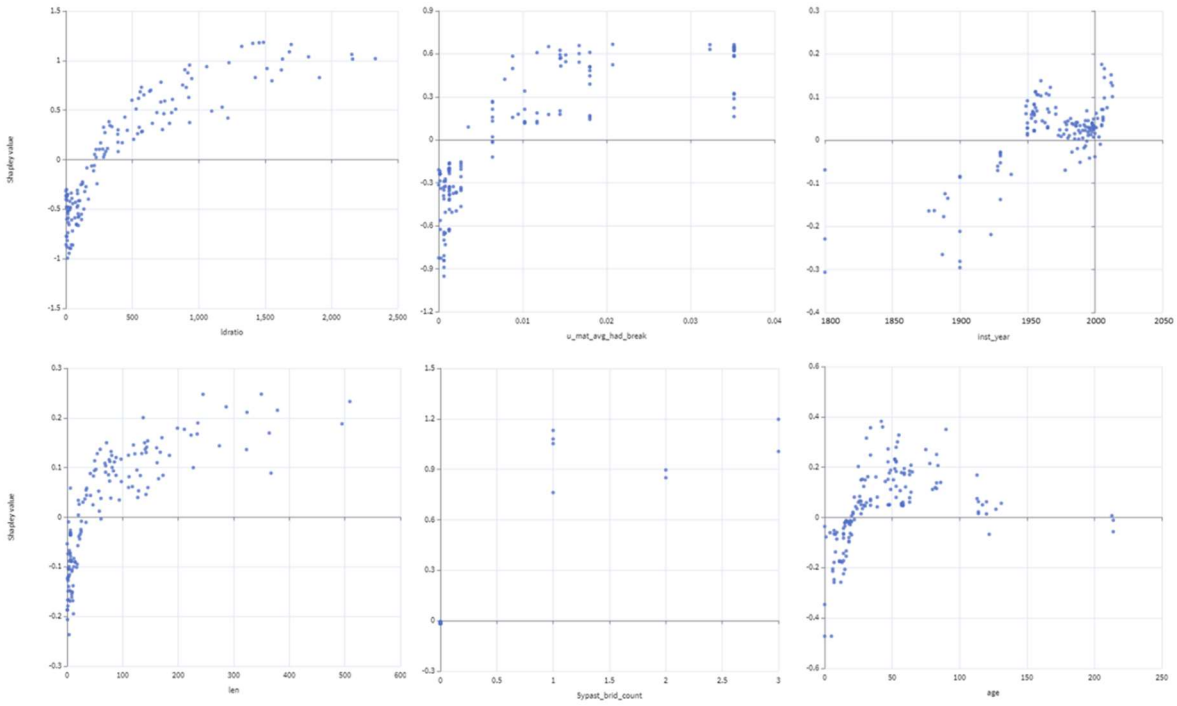


Figure 54: Plots of the contribution by features of greatest impact, as measured by Shapley values on a random selection of samples from the test set.

5.4.2.2 Feature Partial Dependence

Another means of understanding the impact is via partial dependence plots. Partial dependence plots simulate measuring the direct impact of feature values on predictions by re-running the model inference for part or all the test sets with modified values for the feature of interest. For example, a partial dependence test on the five-year break history would run the full test set through model inference, first testing it with values of 0 failures for each sample, then with values of 1 failure for each sample, and so on to the limit of the available values. The plots in Figure 55 group numerical features into bins, showing on the x-axis the feature value, and on the y-axis the relative increase or decrease of the log odds of a "true" (i.e., a break) prediction within that bin vs the average log odds of a "break" prediction in the full test set. Also plotted against the secondary (right) y-axis are bar charts of the percentage of samples falling into each bin.

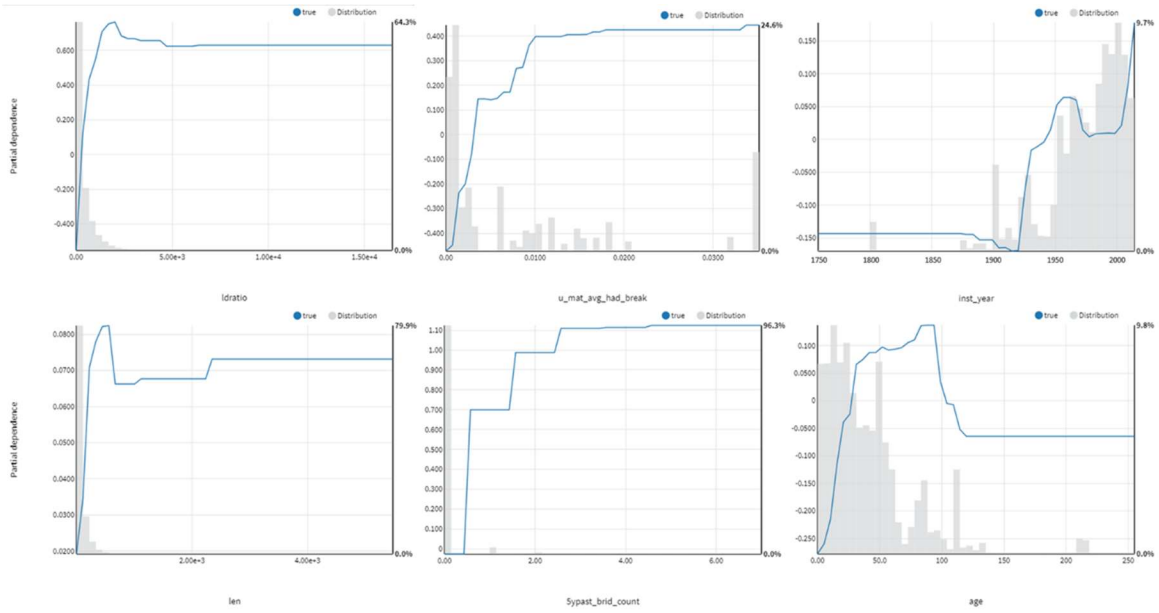


Figure 55: Partial dependence plots for features of interest, together with bar charts of sample density.

Partial dependence plots by age are of particular interest. These represent a potential practical application of the model in the task of projecting the break rates in the longer-term future. These projections would be calculated by taking a given set of pipes and running them through model inference with their ages incremented.

5.5 Alternate Target Variable: Annual Break Rate

While the five-year-forward break rate has been selected as the primary target variable, the selected model achieves comparable performance on the selected metrics when tested with the one-year-forward break rate (i.e., the annual probability of failure).

Table 34: Comparative results predicting one-year and five-year break status with the same model.

	LOGO Results Predicting		
	5-years	1-year	Relative
By ROC AUC	0.864	0.863	99.7%
By Lift at 10%	5.092	5.128	100.7%
By Log Loss	0.117	0.041	285.4%
By F1 Score	0.272	0.123	45.2%

Performance is not meaningfully different on the primary performance metrics of AUC and Lift at 10%, indicating that the model ranks pipes based on their relative one-year or five-year failure risk equally well. Unsurprisingly, the Log Loss and F1 Score metrics are substantially different. Shortening the time window for the target variable has decreased the percentage of positive instances from 3.47% to 0.86%, thereby impacting these metrics. Both of these metrics are impacted by the degree of class imbalance in the test data, making them inappropriate for comparison of different time windows.

5.6 Chapter Summary

This chapter described the results obtained from the generalized machine learning model for pipe failure prediction, following the methodology described in Chapter 4.5. The Feature Preprocessing layer was shown to improve performance for models that were tested on data from a utility which was excluded from the training data. Results from the Failure Classification layer simulate a new utility using a pretrained model created using data from other utilities, which is the intended practical application of this research. The resulting model slightly outperformed the results of training and testing separate models on each utility in isolation, and performed better than 90% as well as a model trained on pooled data from all utilities including the utility being tested. Results from the final Calibration layer demonstrated a strong ability to predict the number of breaks in a cohort, achieving an R^2 value of 0.99 when comparing the predicted and actual number of breaks per cohort in a five-year period.

Chapter 6 provides a discussion of the results obtained by following the methodology described in Chapter 3. It begins with a discussion of the potential implications of results of the exploratory analysis described in Chapter 4, and then provides a detailed discussion of the results of the final generalized machine learning model for pipe failure prediction as provided in this chapter.

Chapter 6
Discussion: Applying Results to Pipe Replacement Decisions

Much of the value from this study comes as a result of the sheer size and variety of the data collected. This has allowed novel insights to be drawn from three different aspects of this study: from the exploratory data analysis, from the exploratory modeling, and from the final generalized model.

6.1 From the Individual Predictive Feature Analysis

This section presents key observations from the exploratory data analysis. While these observations are not the primary focus of this research, they present potential directions for future research.

6.1.1 Segment Length as a Simple Predictor of Break Risk for Distribution Mains

It is highly intuitive that segment length would be a predictor of break risk. A longer segment simply introduces more points for potential failure. Furthermore, segments are logical groupings of pipe sticks, rather than genuinely discrete physical things. Changes to how these groupings are formed should not change the total aggregate number of failures. So, using smaller groupings to define the segments (and therefore more segments) should result in fewer breaks per segment. Were this not the case, there would need to be a confounding variable that impacts both the break rate and the segment length.

This intuition has been confirmed with the analysis of data from this study. The linear relationship between segment length and break rate is clear up to a segment length of 500m. This length limit includes 99% of the segment-year records in this data set and will include all of the distribution mains.

This finding confirms the soundness of the common industry practice of reporting break rates on a unit length basis (typically per 100 miles per year in the US, and per 100km per year elsewhere).

The relationship between failure risk and length for long diameter transmission mains is less clear from the data in this study. This is left as an area for future research.

6.1.2 Length to Diameter Ratio as a Simple Predictor of Break Risk

Pipe diameter was a clear predictor of pipe break risk in the data from this study, with smaller diameter pipes failing at higher rates. The relationship was consistently present across all materials,

diameters, and utilities. Prior research had indicated higher rates of breaks for small diameter mains for individual utilities and material types. The consistency of the relationship in this large and diverse dataset establishes the relationship more clearly than ever before.

When normalizing the break rate by pipe length (i.e., breaks per 100 km per year), the impact of diameter became even more clear and consistent. The relationship appears to be inverse linear, such that the length/diameter ratio offers a consistent linear relationship with the rate of breaks per year.

This inverse linear relationship between diameter and break rate, as shown in Chapter 4.2.1, is a novel finding. It is of interest for two reasons. First, it offers an alternative normalization for the break rate. Second, it suggests a potential area of improvement for pipe design specifications.

While Chapter 4.1.3 confirmed the appropriateness of normalizing break rates for length (breaks per 100 km per year), this finding suggests that an additional normalization by diameter may be useful. Mathematically, this would be:

$$BreakRate = \frac{Breaks * Diameter}{Length * Years} \quad \textbf{(34)}$$

This equation can be rearranged to illustrate a simplified description of the normalization:

$$BreakRate = \frac{Breaks}{\frac{Length}{Diameter} * Years} \quad \textbf{(35)}$$

This rearrangement highlights the ratio of pipe length to pipe diameter, allowing the pipe's length to be expressed in pipe diameters. For example, a 1km long pipe with a 1,000mm diameter (1m diameter) would be 1,000 pipe diameters long, whereas a 1km long pipe with a 100mm diameter (0.1 m diameter) would be 10,000 pipe diameters long. A suggested scale is breaks per million pipe diameters of length per year. A normalization of the break rate per million diameter-years is analogous to the break rate per 100 km-years already broadly in use and is identical for the very common pipe diameter of 100mm. Testing on our dataset shows that this scaling yields results on a similar order of magnitude on our dataset, with most values falling in an easily human readable range

of 1 to 100. A further benefit of this scale is that Length / Diameter is unitless, so this normalization would be equally intuitive for both metric and imperial unit users.

The potential area of improvement for pipe design specifications requires an examination of the current standard design specifications. Standards for pipe design require that for a given pressure class, thicker walled pipe is to be used for larger diameters. This standard is illustrated in Table 35, which shows the North American standard for minimum wall thickness of different diameters of pipe and different pressure classes. For any given pressure class in the table, the required wall thickness increases as the pipe diameter increases.

Table 35: North American standards for minimum wall thickness by pressure class and pipe diameter. From *AWWA C151/A.21.51 Ductile-Iron Pipe, Centrifugally Cast*.

Nominal Size (inches)	Outside Diameter (inches)	Pressure Class				
		150	200	250	300	350
		Wall Thickness (inches)				
3	3.96	0.25	0.25	0.25	0.25	0.25
4	4.80	0.26	0.26	0.26	0.26	0.26
6	6.90	0.25	0.25	0.25	0.25	0.25
8	9.05	0.27	0.27	0.27	0.27	0.27
10	11.10	0.29	0.29	0.29	0.29	0.29
12	13.20	0.31	0.31	0.31	0.31	0.31
14	15.30	0.33	0.33	0.33	0.33	0.33
16	17.40	0.34	0.34	0.34	0.34	0.34
18	19.50	0.35	0.35	0.35	0.35	0.35
20	21.60	0.36	0.36	0.36	0.36	0.39
24	25.80	0.38	0.38	0.38	0.41	0.44
30	32.00	0.39	0.39	0.43	0.47	0.51
36	38.30	0.43	0.43	0.48	0.53	0.58
42	44.50	0.47	0.47	0.53	0.59	0.65
48	50.80	0.51	0.51	0.58	0.65	0.72
54	57.56	0.57	0.57	0.65	0.73	0.81

The rationale behind this standard comes from Barlow's Formula for thin-walled pipe (Adams et al., 2018) and the Lamé equations for thick-walled pipe (Capecchi & Ruta, 2015).

Barlow's Formula, originally pushed in 1836 and shown in (36), relates the Allowable Stress (also referred to as the Hoop Stress), outside diameter, wall thickness, and internal pressure.

$$P = 2St/D \tag{36}$$

Where P = Pressure inside a pipe

S = Hoop Stress, an empirically calibrated Allowable Stress value per material

t = Wall thickness

D = Outside pipe diameter

The Lamé equation offers a similar relationship between pressure and hoop stress for thicker walled material, as shown in (37).

$$S = \frac{P_{int}a^2 - P_{ext}b^2}{b^2 - a^2} + \frac{a^2b^2(P_{int} - P_{ext})}{(b^2 - a^2)r^2} \tag{37}$$

Where P_{int} = Internal pressure

P_{ext} = External pressure

S = Hoop stress

a = Radius to inside of pipe wall

b = Radius to outside of pipe wall

r = Point of interest on the pipe wall

Two simplifying assumptions allow us to express the Lamé equation in a simplified manner. First, assuming that the external pressure is zero, and second assuming that we are interested in the point on the pipe wall where stress is maximized, which is where r = a. In this case, the equation can be simplified as follows:

$$S = \frac{P_{int}a^2}{b^2 - a^2} + \frac{a^2b^2P_{int}}{(b^2 - a^2)a^2} \tag{38}$$

$$S = \frac{P_{int}a^2}{b^2 - a^2} + \frac{b^2 P_{int}}{b^2 - a^2}$$

(39)

$$S = P_{int}\frac{a^2 + b^2}{b^2 - a^2}$$

(40)

$$P_{int} = S\frac{b^2 - a^2}{a^2 + b^2}$$

(41)

Once rearranged as shown in (41), it is apparent that the numerator ($b^2 - a^2$) increases as the wall thickness ($b - a$) increases, whereas the denominator ($a^2 + b^2$) increases as the diameter increases. Hence for both the thin-walled and thick-walled equations, at a given pressure and allowable hoop stress (which is specific to the material type), these equations state that the required wall thickness increases as the diameter increases.

While the rationale is clear, the strong relationship found between diameter and break rate suggests that the design standards may not be achieving the desired outcome of preventing failures across all diameters. There are two plausible reasons why this may be the case.

One possible reason emerged from a study of the origins of the Barlow equation. This study noted flaws in the original 1836 derivation of the equation, the results of which being that "Present design practice is over-conservative for thick wall pipe, and potentially unconservative for thin wall." (Adams et al., 2018) This agrees with the observations in this study that small diameter pipe (with thinner required wall thicknesses) consistently fails at a higher rate than large diameter pipe (with thicker required wall thicknesses).

A second possible reason comes from Makar et al. (2001), which observed that as diameters decrease in cast iron pipe, the most common failure mode switches from longitudinal cracking to circumferential cracking. A longitudinal crack is the expected result of the internal pressure exceeding the hoop strength of the pipe. This is the failure mechanism considered in Barlow's Formula and the Lamé equation and, together with crushing forces, the failure mode considered in pipe design specifications. The author has personally observed this failure mode during burst testing

of pipes. Circumferential cracking failure, on the other hand, can be caused by either shear or bending forces, such as those caused by frost heave (Selvadurai & Shinde, 1993). These prior results, taken together with this new observation of the inverse relationship between diameter and failure rate, suggest that these failure modes have not been adequately considered in the design specifications for small diameter pipe. Exploring the implications of these alternative failure modes on pipe specifications is a recommended area for further future research.

Furthermore, the inverse linear relationship between diameter and failure rates per unit distance opens the potential for a new normalization method for reporting break rates. Presenting the break rate on the basis of Breaks per Million Pipe Diameters of Length offers a more consistent and stable normalization of the break rate.

A potential explanation for the relationship between diameter and break rate comes from pipe design standards. Pipes are designed based primarily on burst strength equations. These equations describe the internal pressure at which a pipe will suffer a longitudinal split. These equations call for thicker pipe walls at larger diameters for a given internal pressure. However, prior research has shown that at smaller diameters the principal failure mode ceases to be longitudinal splits and becomes instead circumferential splits.

The combination of dramatically increasing failure rates at small diameters, and a change in the failure mode at small diameters away from the mode used to create design specifications offers an important possibility. It may be that bending and/or shear forces acting on the pipeline are the primary stresses causing failures in small diameter mains. This hypothesis could be confirmed by modeling the risk of failure due to such forces of varying magnitudes at different diameters and the accompanying standard wall thicknesses and comparing the resultant expected failure rates to those observed in this study. This is left as an area for further research.

It is worth noting that if this proves to be the case, there could be substantial implications for industry. It would mean that a substantial fraction (perhaps more than 50%) of water pipeline failures could be avoided through the introduction of new pipeline design standards for small diameter pipes. The value of this line of research and standards revision was also highlighted by Adams et al. based on their review of the origin of the Barlow equations, stating, “The industry should therefore consider revising OCTG burst ratings and accompanying design practice to achieve a more uniform safety

level over the full D/t range of casing and tubing" (Adams et al., 2018, p. 1). Globally, such a change could translate to tens of billions of dollars per year worth of avoided damage.

Even without modification to pipe design standards, the strength of the relationship between diameter and failure rate may offer opportunities for efficiency gains at utilities. Simply increasing the standard diameter used for distribution mains (often 100mm or 4-inch in North America, smaller in Europe) could economically reduce failure rates. As previously observed (Neelakantan et al., 2008), assuming a lower break rates for larger diameters would result in longer optimal replacement periods and in some cases lower total costs if larger diameters were used than was necessary for the network capacity, despite the increased material and construction costs. Based on the results of this study, changing from a standard of 100mm to 150mm could reduce main breaks in these mains by roughly 1/3, and changing from 100mm to 200mm could reduce main breaks by roughly half. A further potential benefit of larger diameters is improved energy efficiency in networks (Dziedzic & Karney, 2014). A comprehensive review of optimal diameters, considering the operational cost reductions from both reduced energy costs and reduced break repair costs, is recommended as an area for further research.

6.1.3 Age as a Complex Predictor of Break Risk

As with pipe length, pipe age is an intuitive predictor of break rates. The concept of the bathtub curve of asset life is an established concept that is well accepted by most engineers.

Surprisingly, the data in this study did not conform to these expectations for most materials. The relationship between failure rate and age in the data from this study was more complex. As was found in prior studies, failure rates began low and then increased for an initial period. This period of rising break rates varied by material and utility, with a typical value of 30 to 40 years. A period of declining break rates followed, also with a varying duration. A second rise in the break rate would follow for materials with sufficient history. Beyond that point the data was too sparse to determine whether another fall in the break rate occurs or if the rise is indefinite.

The implications for industry of the rise and then fall of the failure rate are substantial. One practice used for selecting pipes to replace is to track the failure rates of each cohort, to project them forward linearly, and select the cohorts with the highest future expected failure rates. A second practice is to track the cumulate failure rate over time for a given cohort and use this rate as a prioritization method. The presence of a rise and subsequent fall of failure rates with age would mean

that either practice runs a risk of selecting pipes for replacement just as they are entering a long period (in some cases over 50 years) of low failure risk. Confirming this pattern, and making industry professionals aware of it, could prevent tens of billions of dollars annually in unnecessary capital expenditures around the world.

While the pattern is not intuitive, an alternate paradigm of pipeline failures could explain it. Rather than consider "age" as the cause of failure, consider instead that there will be a small number of potential time-dependent processes which can lead to failure. Consider a cast iron pipeline with a bituminous external coating. Certain points along the pipeline will have a thin or damaged coating and will be subject to external corrosion. The corrosion will progress at a speed described by some random distribution. These points will reach the point of failure at another random distribution, determined by the original wall thickness and safety factor as well as the random distribution of the corrosion speeds. If the number of potential points of failure for this particular mechanism is finite, then this would form a "wave" of failures with a characteristic shape, including a peak and then a valley. Once the bulk of this "wave" had passed, nearly all the points along the pipeline subject to that degradation mechanism would have failed, with the remaining points simply not having the appropriate conditions for the mechanism to take place. By contrast, a degradation process which applies to all points along the pipeline could show indefinite growth in the failure rate until each stick of the pipeline had failed and been replaced. Embrittlement of PVC with age is an example of such a process. While exposure to sunlight speeds up this process, it will proceed regardless of environmental conditions.

Under such a paradigm, any given cohort of pipe would be subject to one or more waves of failure caused by specific degradation mechanisms that applied only to certain sticks of pipe. Eventually a ubiquitous failure mechanism would take hold and cause an indefinite rise in the break rate; however, it is possible that the time until this occurs has been substantially underestimated for some materials.

Further investigation of this hypothesis would require a substantial program of material testing, modeling, controlled experimentation, and likely the examination of many samples of exhumed pipe. This is left as a direction for future research.

6.2 From the Exploratory Models

A key observation from the exploratory models was the effectiveness of decision tree-based models in forecasting failure risk. Decision tree-based models, including both simple decision tree and more

complex Random Forests and Gradient Boosted Trees, consistently performed well in this study. This finding was true for both problem formulations studied: predicting whether a break will occur in a segment-year tuple and for forecasting the number of breaks in a cohort-year tuple.

This finding aligns with the results of the exploratory analysis. Many of the powerful predictive variables, such as diameter, had non-linear relationships with the break rate. Others, such as age, had non-monotonic relationships, wherein the break rate increased in certain ranges but decreased in others. Tree-based models are effective at representing these types of complex relationships in numerical data.

The performance of the exploratory models in the segment-year break prediction was noteworthy from a perspective of practical application to industry. Their performance against the metric of cumulative lift at 10% highlights this applicability. The baseline models were generally able to achieve a lift at 10% of roughly six times. This means that a utility using this method to select 10% of their pipes for replacement or rehabilitation would be able to avoid 60% of the breaks in the upcoming year. This represents a highly efficient rehabilitation / replacement program using quite a simple model.

These models also trained quickly, operate quickly at inference time, and are relatively explainable. The strong performance of these models in the domain leads to a clear recommendation for their use in modeling pipe failure risk.

The caveat is that these models require access to a large set (both in number of pipes and failure history) of clean data from each utility and would furthermore require each utility to train their own machine learning model.

6.3 From the Generalized Machine Learning Model

This section presents observations and discussion of the results from the final generalized machine learning model.

6.3.1 The Case for Predicting Probability of Failure in a Given Time Period

This study has demonstrated that the Probability of Failure for a given pipe in a given time period $\text{PoF}(t \mid \boldsymbol{x})$ can be estimated by a pretrained machine learning model. This is the recommended target

variable for the application of machine learning to pipe failure prediction due to the broad practical applicability of the results, both to utility decision making and engineering practice.

In survival analysis, as described in Chapter 2.2.3, one of the core measures used is the hazard function. This can be expressed as a conditional hazard function for a given member of the population $h(t \mid \boldsymbol{x})$, or as an aggregate hazard function for a cohort $h(t)$. The conditional hazard function $h(t \mid \boldsymbol{x})$ is the probability of a member of the population described by feature vector x to experience a failure during time period t, conditional upon having survived to the start of time period t. This is precisely what is provided by a machine learning model in estimating $PoF(t \mid \boldsymbol{x})$. All other common metrics for survival analysis, such as the failure density function, the cumulative failure function, the survival function, and mean time to failure, can be calculated by straightforward numerical methods using the conditional hazard function. As such, the estimate of $PoF(t \mid \boldsymbol{x})$ is a target variable appropriate for use in survival analysis.

In failure risk assessment, as described in Chapter 2.2.4, the total risk of failure of an asset in a time period t is described as the product of the probability of failure and consequence of failure. The probability of failure of a given asset in a given time period is precisely what is provided by a machine learning model in estimating $PoF(t \mid \boldsymbol{x})$. As such, the estimate of $PoF(t \mid \boldsymbol{x})$ is a target variable appropriate for use in failure risk assessment, when combined with a separate method of estimating the consequence of failure.

In decision optimization, as described in Chapter 2.2.5, calculations are made to determine choices which result in the maximum benefits or minimum costs. A generalized cost equation for pipe replacement decisions includes the probability of failure in a given time period for an asset with specific features (length, diameter, material). This is precisely what is provided by a machine learning model in estimating $PoF(t \mid \boldsymbol{x})$. This can be used in optimization calculations, simulating different design decisions by changing features of $\boldsymbol{x}$ (e.g., different materials, diameters, etc.) and searching for the optimal decisions. By varying the value of t (and adjusting any features in $\boldsymbol{x}$ as necessary), long term cost estimates can also be calculated in support of these optimization calculations. As such, the estimate of $PoF(t \mid \boldsymbol{x})$ is a target variable appropriate for use in decision optimization.

In utility capital improvement planning, as described in Chapter 2.1.3.2, utilities seek to make pipe replacement decisions which will improve their performance against key performance indicators.

Several common performance indicators, such as water loss rate and breaks per 100km per year, can be directly calculated using the probability of failure of a given pipe in given time period. This is precisely the output provided by a machine learning model estimating $PoF(t \mid \boldsymbol{x})$. This can be used to prioritize pipes for replacement in a manner that will improve these metrics in a predictable manner. As such, the estimate of $PoF(t \mid \boldsymbol{x})$ is a target variable appropriate for use in utility capital improvement planning.

In utility master plans, as described in Chapter 2.1.3.1, long term projections of capital and operational costs are required. A significant operational cost element is the costs associated with water main breaks. Estimating the future number of water main breaks can help plan for these costs. When calibrated linearly to provided Expected Number of Failures, this quantity can be calculated using a machine learning model's estimate of $PoF(t \mid \boldsymbol{x})$. Furthermore, master planning involves analyzing the tradeoff between capital costs (cost of replacement) and operational costs (cost of responding to failures) associated with various water main replacement program decisions. The estimates of Expected Number of Failures under different simulated scenarios can be provided by calculating $PoF(t \mid \boldsymbol{x})$ with modifications to the values of t and $\boldsymbol{x}$ to match these scenarios. As such, the estimate of $PoF(t \mid \boldsymbol{x})$ is a target variable appropriate for use in utility master planning.

This choice of probability of failure of a given pipe in a given time period as a target variable is appropriate for all of the applications discussed in this thesis. As a highly practical choice for application of results, it is recommended for use.

6.3.2 The Case for AUC and Lift at 10% as Standard Performance Metrics

A wide range of performance metrics are available for machine learning classifiers. Table 36 notes several desirable criteria for a performance metric and which of the available performance metrics meet each criterion. The criteria considered are:

- **Is a standard metric**. A standard machine learning performance metric is desirable: one that is well understood, clearly documented, and ideally available in standard software packages. This ensures that the metric itself is calculated consistently to allow comparison across studies. Custom metrics, requiring manual implementation for each study, introduce risk of a metric that is nominally the same being implemented differently.

- **Independent of decision threshold**. Most applications of pipe failure prediction rely on the relative ranking of pipes. A metric which depends on the selection of a decision threshold does not facilitate comparison of the fitness for use of different models.
- **Can be applied to all pipes**. Certain metrics can only be applied to pipes with recorded history of failure. This limits their broad applicability and understandability.
- **Shows ability to rank pipes**. Certain metrics demonstrate the ability of a model to rank the relative failure risk of all pipes. This is desirable for long-term planning.
- **Focus on high-risk pipes**. Certain metrics focus on the ability of a model to successfully identify the highest risk pipes. This is desirable for medium-term replacement decisions.
- **Simple to explain**. A metric with a simple explanation that will be intuitive for a layperson is far more likely both to be considered in practice and less likely to be misapplied due to incorrect interpretation.

Table 36: Potential performance metrics vs desirable criteria.

	Is a standard metric	Independent of decision threshold	Can be applied to all pipes	Shows ability to rank pipes	Focus on high-risk pipes	Simple to explain
Accuracy	y		y			y
Precision	y		y			y
Recall	y		y		y	y
F1 Score	y		y			
Maximum F1 Score	y	y				
Area under ROC curve	y	y	y	y		y
Lift at 10%	y	y	y		y	y
Log loss	y	y	y			
Concordance index	y	y		y		
Area under break capture curve		y	y	y		y

Many standard metrics, such as Accuracy, Precision, Recall, and F1 Score, can be immediately discounted for recommendation, as they rely on the decision threshold.

Log loss and Maximum F1 Score are both useful measures. Neither, however, is particularly focused on the relative ranking of pipes nor the performance on high-risk pipes. Most importantly, neither has a simple and intuitive explanation. Further, neither has an explanation in the domain of pipe failure prediction which is intuitive and simple to explain to a layperson.

The concordance index is an effective measure of the ability of a model to rank pipes. It is, however, only applicable to pipes with at least one failure recorded in their history. Furthermore, it lacks an intuitive explanation.

Area under the break capture curve is a well-crafted metric for this application. It has a simple and intuitive explanation and would likely be correctly used to select models for pipe replacement prioritization. Its only downside is that it is not a standard metric and must be separately coded for each study. This introduces a substantial risk of variations in understanding or implementation, which would result in comparison of different measures that are nominally the same.

Area under the ROC curve and Lift at 10% each meet nearly all the desired criteria. Each is a standard metric already implemented in most machine learning packages. Each has an intuitive explanation in the domain of pipe failure prediction. The area under the ROC curve measures how well the model ranks the pipes in order of risk of failure, with a score of 0.5 indicating random ranking and 1.0 indicating perfect ranking. The Lift at 10% measures what portion of breaks could be avoided if the 10% most at-risk pipes were to be replaced. When considered jointly, the two can confirm that a model is broadly effective at ranking pipes and particularly effective at identifying the highest risk pipes. This makes them appropriate metrics for comparing models.

It is further noted that in this study, both metrics were stable regardless of whether the model was used to predict whether a break would occur in a one-year or five-year period. If this finding is consistent in other studies, and in particular, if the same is true over longer time periods such as 10 or 20 years, it would mean that performance by these metrics could be compared across studies using a range of different time periods. This is left as an area for future research.

6.3.3 Effectiveness of Preprocessing Layer

Ubiquitous in the literature is the claim that substantial manual data cleansing is required on utility data before application of machine learning for pipe failure prediction (Delnaz et al., 2023). This study has demonstrated that automated preprocessing can remove the need for this cleansing.

A summary of comparable results from the literature (see Chapter 2.3.2.2 for details) in which AUC was reported is provided in Table 6 . Performance as measured by AUC ranges from 0.81 to 0.88 when a time split is applied between train and test sets and between 0.80 and 0.90 when no time separation is enforced. This finding aligns with the results of the exploratory modeling conducted in this study and described in Chapter 4.3.1.6. This study showed that an AUC score of 0.87 can be achieved (see Chapter 5.2.3) with training and testing on each utility separately, using only the automated preprocessing steps described here, and no manual or utility-specific data cleansing rules. This result nearly matches the top results reported in the literature of 0.88 when a time split is enforced between train and test data sets.

6.3.4 Benefit of Pooling Data

In all tests conducted during this study, the Inclusive data scheme outperformed the Isolated data inclusion scheme by every performance measure (see Chapter 5.2). The average performance difference across the metrics was approximately 10%. This result suggests that even utilities with strong record keeping and data management practices would benefit from pooling data for the training of machine learning models. Performance on the Inclusive data selection scheme reached an AUC score of 0.89 on the preprocessed data, exceeding the highest score found in **Error! Reference source not found.** in which a time split is applied between the train and test sets. This result suggests that the benefits of pooling data exceed the benefits of investing in extensive data management activities.

On the converse, the administrative burden for utilities to share their data is not trivial. This fact was demonstrated by the time required to obtain data in this study and the small proportion of contacted utilities who were willing to share data for this study, as described in Chapter 3.2. Whether a 10% improvement in the efficiency of their capital improvement programs is a sufficient benefit to entice utilities to overcome the administrative obstacles to sharing data is an area left for future research.

6.3.5 Practicality of Using a Pretrained Model

The ability of a pretrained model to extrapolate both forward in time and to new utilities concurrently has been demonstrated. The performance reported in Chapter 5.2.3 shows that an AUC score over 0.86, and a Lift at 10% of over 5x, is feasible when using a model trained using only historical data

from other utilities, and without any utility-specific data cleansing. This result essentially matches the performance obtained when training and testing models on each utility in isolation during this study, confirming the findings in Daulat et al. (2024) that pooled data from other utilities can match the performance of a utility using its own data. Furthermore, it is comparable to the highest performance found in the literature of 0.88 (see **Error! Reference source not found.**).

Consequently, the use of a pretrained model using only a small number of commonly available features can perform as well as a model specifically engineered for an individual utility using all available data and manual data cleansing. The most critical features are length, material, diameter, installation date, and the recent failure history (within the past five years), which should be available at nearly all utilities.

This finding offers significant opportunities for operational efficiency among utilities, removing the need for each to train their own models. Furthermore, it opens up the practical possibility for small utilities, who may lack the capacity for extensive data management programs or the ability to train and test their own models to nevertheless benefit from the application of machine learning.

6.3.6 Areas for Model Improvement

Several areas were identified in which stronger performance may be desirable.

First and foremost, the model performance was worse on pipes which had experienced previous failures. This is likely a result of the fact that these make up a small proportion of the dataset. Stronger performance on these high-risk pipes would be desirable in a model. Training a separate model specifically on these pipes may prove beneficial and is left as an area for future research.

Second and related, the model performance was weaker on pipes with a recorded history of rehabilitation. This finding may indicate that further information, such as the type of rehabilitation or the time since rehabilitation, could benefit the model. Investigation into effective features to improve performance on this subpopulation is left as an area for future research.

6.4 Chapter Summary

This chapter provided a discussion of the results obtained in this research project. Two interesting findings from the exploratory analysis were discussed. The strong inverse relationship between pipe diameter and failure rate has possible implications for both pipe design standards and for the selection

of pipe diameters. The observed relationship between pipe age and failure rate, with peaks and subsequent declines, is a novel finding which may have significant implications for pipeline management decisions. The practical applicability of the generalized machine learning model for pipe failure prediction was discussed in detail. The preprocessing layer was effective in allowing a pretrained model to be applied to new utilities, and has the potential for application to other problem domains where pretrained models would offer practical value. Pooling training data among utilities was also shown to offer the potential for a roughly 10% improvement in the effectiveness of a pipe replacement program at reducing pipe failures. Several areas for model improvement were also provided.

Chapter 7 provides the final conclusions of this research, including the limitations of the methods applied and potential areas for future research.

Chapter 7
Conclusions

7.1 Key Findings and Potential Impact

This thesis has demonstrated that a practical machine learning model for prediction of water pipe failure is indeed possible. A method has been presented for integrating data from multiple utilities, without the need for manual data cleansing and estimating both the calibrated Probability of Failure and the Expected Number of Failures within a given timeframe. The method has been demonstrated on data from six water utilities across four countries. It was found that pooled data from multiple utilities yields models that outperform those trained on a single utility in isolation. With the application of the preprocessing and calibration steps proposed, it has been shown that models trained only on other utilities' data can match the performance of models trained on a single utility in isolation. The model presented achieved an AUC of over 0.86, and Lift at 10% of over 5x. This performance, done without using data from the target utility for training or the application of manual data cleansing, is comparable to that reported in the literature for models trained on individual utilities' data with manual preprocessing applied. The expected number of failures, when aggregated to cohorts, yielded correlations with R^2 value of 0.99 in comparison to actual future failure numbers.

7.1.1 Key Contributions to the Research

The three most significant contributions to the research are the creation of a large pipe failure dataset from diverse sources, the demonstration of generalizability of pretrained models across utilities, and the cross-encoding method for handling of categorical features.

The dataset itself is of significant value, as prior research into pipe failure prediction has generally used data from a single source (one utility, or one country). The large and diverse dataset enables generalized observations which apply across a range of geographic and environmental regions.

The demonstration that a model trained using data from a group of utilities can be effectively used by other utilities that did not contribute to the training data is also significant. Prior published research has demonstrated that a model trained on data from one utility can predict future pipe failures in that same utility. This is of limited practical applicability, as most utilities lack the data and expertise to train their own model. The novel finding of generalizability from this study shows that pipes around the world fail under consistent enough patterns that a more general model can be

created. This confirms the feasibility of a more practical approach to research into machine learning for pipe failure prediction, searching for broadly applicable models and results, rather than ones specific to one utility or country.

The proposed method of categorical feature cross-encoding offers a valuable technique to support such generalized models. This method overcomes two significant limitations of dummy encoding, the conventional approach for handling categorical variables in machine learning models. The first limitation is that new categories not seen in the training data cannot be used at test time when dummy encoding is used. By representing each category by a vector of statistics drawn from samples with that label, new labels not seen at training time can now be used. The second limitation is that dummy encoding can produce a large number of features: one for every value of the categorical variable. Cross-encoding offers a simple means of limiting the number of features created. These effects together may make cross-encoding a viable alternative to dummy encoding for a wide variety of situations, particularly when data is integrated from disparate sources.

7.1.2 Findings and Potential Impact

An argument has been made for the preferred target variable for pipe failure prediction using machine learning. Predicting a calibrated probability of failure for a given time period yields results which can be easily applied to several engineering applications, including failure risk analysis (considered jointly with the consequence of failure), survival analysis (as an estimate of the hazard rate, which can be numerically integrated to provide survival and failure curves), and optimization calculations. It can be directly utilized to prioritize pipe replacement in the five-year period typically used for utility capital improvement plans. By linearly scaling to the Expected Number of Failures, total expected failure forecasts can be made per cohort or per utility to support long term budgetary planning.

A further argument has been made for the preferred performance metrics for pipe failure prediction using machine learning. The Area Under the Curve for the Receiver Operating Characteristic is a standard metric which aligns well with the need for relative prioritization of pipes for replacement. Lift at 10% offers a second standard metric with an interpretation that is both intuitive and relevant: if a utility were to select 10% of their pipes for replacement, it shows the fraction of the breaks which could be avoided.

Analysis of such a large and diverse dataset allowed novel observations to be made.

First, it was observed that pipe diameter is strongly related to the failure rate per km, with small diameter pipes failing more than 100 times as frequently as large diameter pipes. It was noted that the correlation is inverse, suggesting that the length to diameter ratio may be a useful normalization factor for failure rates, with failures per million diameters per year on a similar scale to the existing failures per 100 km per year standard. It was further noted that the literature includes observations that the common failure mode in small diameter pipe (circumferential splits) is not accounted for in the common pipe design standards. Combining these observations with the data in this study suggests that changing pipe design standards and diameter selection practices could reduce water main failure rates significantly, perhaps by as much as half. Investigation into these phenomena and opportunities is left as an area of future research.

Second, it was observed that for many materials, failure rates by age do not increase in an exponential manner as is assumed by many models, nor even monotonically. In fact, they tend to reach a peak (between 20 and 50 years, depending on the material), then decline for an extended period, and then eventually rise again. This observation has significant implications for survival curve modeling, the design life and time to failure concepts, and for pipe rehabilitation / replacement programs. One possible explanation is that the first few decades after a pipe is installed represent a secondary wear-in period, where local flaws in material or installation gradually lead to failure, and that genuine wear-out failures do not begin until much later than was previously thought. Investigation into this phenomenon, its causes, and its implications is left as an area for future research.

Analysis of the model results confirmed that even large utilities with extensive records and clean data could benefit from pooling their data to train machine learning models for probability of failure. Furthermore, utilities with limited historical records and less clean data can benefit from a model trained on data from other utilities, reaching performance on par with large utilities training on their own data alone.

The method for integrating data from multiple contributing sources has been demonstrated to be highly effective in allowing results to generalize and extrapolate to new utilities. Central to this method was the concept of encoding categorical features as vectors, drawn from aggregate statistics of numerical features from samples sharing the same label for that categorical feature. This method may be broadly applicable, particularly when pooling data from contributors using different language

or jargon for the application of machine learning models. Further investigation into the applicability of this method, as well as testing of its generalizability, are left as areas for future research.

Finally, it was observed that the model performance is weaker on pipes which have previously failed and which have previously been rehabilitated. An area suggested for further research is the study of the future performance of water pipes which have previously failed and/or been rehabilitated. These pipes are often among those considered for replacement, and a better understanding of their expected future performance would be of great value.

7.2 Limitations of Approach

Despite overcoming several major limitations of prior approaches to machine learning for pipe failure prediction, three significant limitations remain. Two of these stem from the available data, and the third is related to the modeling method itself.

First, no sensor data was used by the models. It is a reasonable hypothesis that sensor data from permanent monitoring devices and/or pipeline inspections contains significant useful information regarding the condition of pipes. Upon initiating this research program, the researchers' intent was to collect such data and integrate it into the machine learning model. Unfortunately, no such data was available from the participating utilities.

Second, no environmental data was used by the models. Environmental data can be grouped into two forms: local data which would apply to individual pipes (e.g. type of bedding supporting the pipe or whether it is covered by grass, sidewalk, or a road), and regional data which would apply to an entire utility (e.g. historical temperatures and rainfalls). Local data was not available from the participating utilities. Regional data was not included due to a concern that it would lead to overfitting in the models. With only one value per utility and six participating utilities, such data would not present sufficient unique data points for the models to infer robust relationships.

Third, it was noted in Chapter 5.4.1.6 that model performance was worse among pipes with recorded history of failures or rehabilitation. Such pipes would be of interest in pipe replacement prioritization programs. Improving the model's ability to discriminate among this subpopulation of pipes would enhance its potential for practical use.

7.3 Future Research

Three broad areas for future research have been highlighted by this thesis: deeper research into key findings from the data, addressing the remaining limitations of the model, and a direct comparison between this model and other approaches.

The large dataset used in this research resulted in two findings which bear further research. First was the strong inverse relationship observed between diameter and failure rates. Deeper research into the physical causes of this observed phenomenon is recommended, as is exploration of the implications for pipe design standards and selection of diameters for installation. Second was the relationship observed between age and failure rates, with peaks and subsequent declines. This novel finding suggests three potential paths of research: determining the root cause of the observed pattern, exploring its implications for the validity of prior research and existing standards which assume an exponential increase of the failure rate with time, and revising the practical risk and optimization engineering calculations to account for this relationship. One possible explanation for the root cause of this pattern would be that some failure modes are applicable only at a finite number of points in a pipe network, as outlined in Chapter 6.1.3. Exploration of this hypothesis by formulating a statistical model and testing its fit to failure data may be a valuable path of research.

The limitations of this model could be addressed by future research projects. Addressing the data limitations would require a significant effort to acquire data from utilities containing pipe sensor measurements and environmental measurements. Acquiring such data from a range of utilities is likely to be a multi-year effort, and may require the continuity of work offered by a research lab focused on this area. Addressing the model underperformance on pipes with recorded history of failure or rehabilitation, however, may be practical using data similar to that used in this study. Training a separate model only on pipes with this history may prove effective. Limiting the training and test set to these pipes, and introducing additional features using this history data, may prove to be effective.

As a final direction for future research, the generalizability of this model makes it particularly well suited to direct comparison with other pipe failure prediction approaches. This model is pretrained, requires a very small number of data fields, and does not require manual data cleansing. As such, the test data used by other research programs could easily be tested using this model as well. This would allow direct comparison of the model results with any other model used to predict probability of

failure per segment. The calibration for expected number of failures per segment allows the model to be directly compared with other models which predict the number of failures per cohort. By treating the model's predicted number of failures per segment per year as the hazard function, it can also be compared directly against survival curve estimates. This would allow the results of this model to serve as a standard comparison point for nearly any other model, thereby facilitating performance comparison across results which cannot currently be directly compared.

References

Adams, A. J., Grundy, K. C., Kelly, C. M., Lin, B., & Moore, P. W. (2018). The Barlow equation for tubular burst: A muddled history. *Society of Petroleum Engineers - IADC/SPE Drilling Conference and Exhibition, DC 2018*, *2018-March*. https://doi.org/10.2118/189681-ms

Al-Ali, A. M., Laurent, J., & Dulot, J. P. (2020). Developing deterioration prediction model for the potable water pipes renewal plan – case of Jubail industrial city, KSA. *Desalination and Water Treatment*, *176*. https://doi.org/10.5004/dwt.2020.25539

Al-Barqawi, H., & Zayed, T. (2006). Condition rating model for underground infrastructure sustainable water mains. *Journal of Performance of Constructed Facilities*, *20*(2), 126–135. https://doi.org/10.1061/(asce)0887-3828(2006)20:2(126)

Al-Zahrani, M., Abo-Monasar, A., & Sadiq, R. (2016). Risk-based prioritization of water main failure using fuzzy synthetic evaluation technique. *Journal of Water Supply: Research and Technology - AQUA*, *65*(2). https://doi.org/10.2166/aqua.2015.051

Almheiri, Z., Meguid, M., & Zayed, T. (2020). Intelligent approaches for predicting failure of water mains. *Journal of Pipeline Systems Engineering and Practice*, *11*(4). https://doi.org/10.1061/(asce)ps.1949-1204.0000485

Amini, M. (2021). *Application of machine learning algorithms to the prediction of water main deterioration*. Concordia University.

Amini, M., & Dziedzic, R. (2022). Comparison of machine learning classifiers for predicting water main failure. *Lecture Notes in Civil Engineering*, *250*. https://doi.org/10.1007/978-981-19-1065-4_42

Aslani, B., Mohebbi, S., & Axthelm, H. (2021). Predictive analytics for water main breaks using spatiotemporal data. *Urban Water Journal*, *18*(6). https://doi.org/10.1080/1573062X.2021.1893363

Atherton, D. L., Morton, K., & Mergelas, B. J. (2000). Detecting breaks in prestressing pipe wire. *Journal / American Water Works Association*, *92*(7), 50–56. https://doi.org/10.1002/j.1551-8833.2000.tb08972.x

Baird, G. M. (2011). The epidemic of corrosion, part 1: Examining pipe life. *Journal - American Water Works Association*, *103*(12), 14–21. https://doi.org/10.1002/j.1551-8833.2011.tb11574.x

Balekelayi, N., & Tesfamariam, S. (2021). Operational risk-based decision making for wastewater pipe management. *Journal of Infrastructure Systems*, *27*(1). https://doi.org/10.1061/(asce)is.1943-555x.0000586

Barton, N., Hallett, S. H., Jude, S. R., & Tran, T. H. (2022a). An evolution of statistical pipe failure models for drinking water networks: a targeted review. In *Water Supply* (Vol. 22, Issue 4). https://doi.org/10.2166/ws.2022.019

Barton, N., Hallett, S. H., Jude, S. R., & Tran, T. H. (2022b). Predicting the risk of pipe failure using gradient boosted decision trees and weighted risk analysis. *Npj Clean Water*, *5*(1). https://doi.org/10.1038/s41545-022-00165-2

BCC Research Inc. (2016). *Special research study: Comparison of water main pipe installation lengths and costs in North and South Carolina: Raleigh, Charlotte, and Spartanburg/Greenville*. https://www.uni-bell.org/portals/0/E-NEWS/MediaFiles/bcc_pipe_report_carolinas.pdf

Beig Zali, R., Latifi, M., Javadi, A. A., & Farmani, R. (2024). Semisupervised clustering approach for pipe failure prediction with imbalanced data set. *Journal of Water Resources Planning and Management*, *150*(2), 1–15. https://doi.org/10.1061/jwrmd5.wreng-6263

Bonakdari, H., Ebtehaj, I., & Ladouceur, J. D. (2023). *Machine learning in earth, environmental and planetary sciences: Theoretical and practical applications*. Elsevier. https://doi.org/10.1016/C2022-0-00549-3

Bostan, M., Azimi, A. H., Akhtari, A. A., & Bonakdari, H. (2021). An Implicit Approach for Numerical Simulation of Water Hammer Induced Pressure in a Straight Pipe. *Water Resources Management*, *35*(15). https://doi.org/10.1007/s11269-021-02992-3

Brander, R. (2004). Minimizing failures to PVC water mains. *Plastic Pipes (12)*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1). https://doi.org/10.1023/A:1010933404324

Cai, Z., Dziedzic, R., & Li, S. S. (2023). Efficiency enhancement of leakage detection and localization methods using leakage gradient and most affected sensors. *Canadian Journal of Civil Engineering*, *50*(12). https://doi.org/10.1139/cjce-2023-0176

Cai, Z., Dziedzic, R., & Li, S. S. (2022). Water distribution system leak detection using support

vector machines. *Lecture Notes in Civil Engineering*, *250*. https://doi.org/10.1007/978-981-19-1065-4_41

Capecchi, D., & Ruta, G. (2015). Strength of materials and theory of elasticity in 19th century Italy: A brief account of the history of mechanics of solids and structures. *Advanced Structured Materials*, *52*. https://doi.org/10.1007/978-3-319-05524-4

Chastain-Howley, A. (2005). Transmission main leakage: How to reduce the risk of a catastrophic failure. *IWA Leakage 2005*, 1–7.

Chen, T. Y.-J., Vladeanu, G., Yazdekhasti, S., & Daly, C. M. (2022). Performance evaluation of pipe break machine learning models using datasets from multiple utilities. *Journal of Infrastructure Systems*, *28*(2). https://doi.org/10.1061/(asce)is.1943-555x.0000683

Clark, R. M., Sivaganesan, M., Selvakumar, A., & Sethi, V. (2002). Cost models for water supply distribution systems. *Journal of Water Resources Planning and Management*, *128*(5), 312–321. https://doi.org/10.1061/(asce)0733-9496(2002)128:5(312)

Cody, R. A., & Narasimhan, S. (2020). A field implementation of linear prediction for leak-monitoring in water distribution networks. *Advanced Engineering Informatics*, *45*(May), 101103. https://doi.org/10.1016/j.aei.2020.101103

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2). https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Dąbrowski, W., & Li, F. (2021). Mortar lining as a protective layer for ductile iron pipes. *International Journal of Civil Engineering*, *19*(4). https://doi.org/10.1007/s40999-020-00585-6

Daulat, S., Rokstad, M. M., Bruaset, S., Langeveld, J., & Tscheikner-Gratl, F. (2024). Evaluating the generalizability and transferability of water distribution deterioration models. *Reliability Engineering and System Safety*, *241*. https://doi.org/10.1016/j.ress.2023.109611

Dawood, T., Elwakil, E., Novoa, H. M., & Delgado, J. F. G. (2020a). Artificial intelligence for the modeling of water pipes deterioration mechanisms. In *Automation in Construction* (Vol. 120). https://doi.org/10.1016/j.autcon.2020.103398

Dawood, T., Elwakil, E., Novoa, H. M., & Delgado, J. F. G. (2020b). Water pipe failure prediction and risk models: State-of-the-art review. *Canadian Journal of Civil Engineering*, *47*(10).

https://doi.org/10.1139/cjce-2019-0481

De Marinis, G., Gargano, R., Kapelan, Z., Morley, M. S., Savic, D., & Tricarico, C. (2009). Risk-cost based decision support system for the rehabilitation of water distribution networks. *Proceedings of the 10th Annual Water Distribution Systems Analysis Conference, WDSA 2008*. https://doi.org/10.1061/41024(340)55

Delnaz, A., Nasiri, F., & Li, S. S. (2023). Asset management analytics for urban water mains: A literature review. *Environmental Systems Research*, *12*(1). https://doi.org/10.1186/s40068-023-00287-7

DiLoreto, G., Hara, M., Kmiec, J., Neumann, K., Olson, D., Peterson, C., Pierce, L., Schlaman, J., Shah, A., & Szafran, J. (2020). *The economic benefits of investing in water infrastructure: How a failure to act would affect the US economic recovery*.

Dziedzic, R. M., & Karney, B. W. (2014). Water distribution system performance metrics. *Procedia Engineering*, *89*. https://doi.org/10.1016/j.proeng.2014.11.200

Eisenbeis, P., Le Guffre, P., & Saegrov, S. (2000). Water infrastructure management: An overview of European models and databases. *AWWARF Infrastructure Conference and Exhibition*, 12.

Ellison, R. D., Leighton, J., Faber, N., Haist, Y., & Conner, D. (2018). *AWWA manual M77: condition assessment of water mains* (1st ed.). American Water Works Association.

Erbay, Ö. O., Zarghamee, M. S., & Ojdrovic, R. P. (2007). Failure risk analysis of lined cylinder pipes with broken wires and corroded cylinder. *Pipelines 2007: Advances and Experiences with Trenchless Pipeline Projects - Proceedings of the ASCE International Conference on Pipeline Engineering and Construction*. https://doi.org/10.1061/40934(252)112

Erdogmus, E., Skourup, B. N., & Tadros, M. (2010). Recommendations for design of reinforced concrete pipe. *Journal of Pipeline Systems Engineering and Practice*, *1*(1). https://doi.org/10.1061/(asce)ps.1949-1204.0000039

Fan, X., Wang, X., Zhang, X., & ASCE Xiong (Bill) Yu, P. E. F. (2022). Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors. *Reliability Engineering & System Safety*, *219*, 108185. https://doi.org/10.1016/j.ress.2021.108185

Fan, X., & Yu, X. (2022). An innovative machine learning based framework for water distribution network leakage detection and localization. *Structural Health Monitoring*, *21*(4). https://doi.org/10.1177/14759217211040269

Fan, X., Zhang, X., & Yu, X. B. (2023). Uncertainty quantification of a deep learning model for failure rate prediction of water distribution networks. *Reliability Engineering and System Safety*, *236*. https://doi.org/10.1016/j.ress.2023.109088

Farmani, R., Kakoudakis, K., Behzadian, K., & Butler, D. (2017). Pipe failure prediction in water distribution systems considering static and dynamic factors. *Procedia Engineering*, *186*. https://doi.org/10.1016/j.proeng.2017.03.217

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8). https://doi.org/10.1016/j.patrec.2005.10.010

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, *57*(3). https://doi.org/10.2307/1403797

Folkman, S. (2018). Water main break rates in the USA and Canada: A comprehensive study. *Mechanical and Aerospace Engineering Faculty Publications*, *March*, 1–49. https://digitalcommons.usu.edu/mae_facpub/174

Fracta. (2022). *Water main condition assessment Fracta overview – Bringing AI to infrastructure*.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1). https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Ge, S. (2016). *Development of a numerical model to analyze the condition of prestressed concrete cylinder pipe (PCCP)*. https://vtechworks.lib.vt.edu/handle/10919/82163

Ghirmay, A. M. (2014). *Asbestos cement pipe condition assessment and remaining service life prediction* [Master's thesis, University of Arkansas]. https://scholarworks.uark.edu/cgi/viewcontent.cgi?article=3283&context=etd

Giraldo-González, M. M., & Rodríguez, J. P. (2020). Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks. *Water (Switzerland)*, *12*(4). https://doi.org/10.3390/W12041153

Giustolisi, O., Laucelli, D., & Savic, D. A. (2006). Development of rehabilitation plans for water mains replacement considering risk and cost-benefit assessment. *Civil Engineering and Environmental Systems*, *23*(3). https://doi.org/10.1080/10286600600789375

GM BluePlan Engineering. (2020). *2020 water and wastewater master plan for Region of Peel*. https://www.peelregion.ca/strategicplan/media/appendix-j-costing-methodology.pdf

Gray, M. R., McCandless, T., & Bonds, R. (2009). *AWWA manual M41: Ductile-iron pipe and fittings* (3rd ed.). American Water Works Association.

Groover, M. P. (2019). *Fundamentals of Modern Manufacturing Materials, Processes, and Systems Seventh Edition*. Wiley.

Hu, Y., Wang, D., & Cossitt, K. (2008). Asbestos cement water mains: history, current state, and future planning. *60th Annual Western Canada Water and Wastewater Association Conference*, 1–13.

Hülsmann, T., & Nowack, R. E. (2004). 70 years of experience with PVC pipes. *Proceedings of Plastic Pipes XII*.

Hunt, E. B., Martin, J., & Stone, P. J. (1966). *Experiments in induction*. Academic Press.

Jernigan, W., George Kunkel, A., Cavanaugh, S., Sayers, D., Brian Skeens, V., Jim Siriano, J., Dan Strub, A., & Sturm, R. (2019). *Key Performance Indicators for Non-Revenue Water* (Issue November). www.awwa.org

Jones, C., & Laven, K. (2008). Water loss & leakage in critical, high risk city areas. *American Water Works Association - American Water Works Association Annual Conference and Exposition, ACE 2008*.

Joshi, T. (2012). *An evaluation of large diameter steel water pipelines* [Masters thesis, University of Texas at Arlington]. https://rc.library.uta.edu/uta-ir/bitstream/handle/10106/11651/Joshi_uta_2502M_11931.pdf

Kabir, G., Tesfamariam, S., Francisque, A., & Sadiq, R. (2015). Evaluating risk of water mains

failure using a Bayesian belief network model. *European Journal of Operational Research*, *240*(1). https://doi.org/10.1016/j.ejor.2014.06.033

Kahn, C., Damiani, A., & Ge, S. (2020). Validation of water main failure predictions: A 2-year case study. *AWWA Water Science*, *2*(3). https://doi.org/10.1002/aws2.1179

Karimian, F., Kaddoura, K., Zayed, T., Hawari, A., & Moselhi, O. (2021). Prediction of breaks in municipal drinking water linear assets. *Journal of Pipeline Systems Engineering and Practice*, *12*(1). https://doi.org/10.1061/(asce)ps.1949-1204.0000511

Khozani, Z. S., Bonakdari, H., & Zaji, A. H. (2017). Efficient shear stress distribution detection in circular channels using Extreme Learning Machines and the M5 model tree algorithm. *Urban Water Journal*, *14*(10). https://doi.org/10.1080/1573062X.2017.1325495

Kull, M., & Flach, P. (2015). Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9284*. https://doi.org/10.1007/978-3-319-23528-8_5

Lambert, A. (1994). Accounting for losses: The bursts and background concept. *Water and Environment Journal*. https://doi.org/10.1111/j.1747-6593.1994.tb00913.x

Lambert, A. (2000). What do we know about pressure: leakage relationships in distribution systems? *IWA Conference on Systems Approach to Leakage Control and Water Distribution System Management*, 1–8.

Lambert, A., & Lalonde, A. (2005). Using practical predictions of economic intervention frequency to calculate short-run economic leakage level, with or without pressure management. *Leakage Conference Proceeding*, *Ili*.

Laven, K., & Lambert, A. (2012). What do we know about real losses on transmission mains? *IWA Water Loss 2012*.

Le Gat, Y., & Eisenbeis, P. (2000). Using maintenance records to forecast failures in water networks. *Urban Water*, *2*(3). https://doi.org/10.1016/S1462-0758(00)00057-1

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11). https://doi.org/10.1109/5.726791

Lu, Z. Q. J. (2010). The elements of statistical learning: Data mining, inference, and prediction, 2nd edition. *Journal of the Royal Statistical Society Series A-Statistics in Society*, *173*.

Makar, J. M., Desnoyers, R., & McDonald, S. E. (2001). Failure modes and mechanisms in gray cast iron pipes. *Underground Infrastructure Research*. https://doi.org/10.1201/9781003077480-47

Malm, A., Moberg, F., Rosén, L., & Pettersson, T. J. R. (2015). Cost-benefit analysis and uncertainty analysis of water loss reduction measures: Case study of the Gothenburg drinking water distribution system. *Water Resources Management*, *29*(15). https://doi.org/10.1007/s11269-015-1128-2

Manda, R. (2012). *Performance of prestressed concrete cylinder pipe (PCCP) in water applications* [Master's thesis, University of Texas at Arlington]. https://rc.library.uta.edu/uta-ir/bitstream/handle/10106/24421/Manda_uta_2502M_11946.pdf

Marzouk, M., & Osama, A. (2017). Fuzzy-based methodology for integrated infrastructure asset management. *International Journal of Computational Intelligence Systems*, *10*(1). https://doi.org/10.2991/ijcis.2017.10.1.50

McFarland, M. L., Water, E., & Coordinator, Q. (2012). Drinking water problems: Corrosion. *Texas A&M Agrilife Extension*, *7*(E-616).

Mohammadi, M. M., Najafi, M., Serajiantehrani, R., Kaushal, V., & Hajyalikhani, P. (2021). Using machine learning to predict condition of sewer pipes. *Pipelines 2021: Planning - Proceedings of Sessions of the Pipelines 2021 Conference*. https://doi.org/10.1061/9780784483602.022

Neelakantan, T. R., Suribabu, C. R., & Lingireddy, S. (2008). Optimisation procedure for pipe-sizing with break-repair and replacement economics. *Water SA*, *34*(2). https://doi.org/10.4314/wsa.v34i2.183642

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*. https://doi.org/10.1145/1102351.1102430

Oliveira, I. (2019). *STE - A survival tree ensemble: Application to the prediction of pipe failures in water supply systems* [Master's thesis, Lisbon Technical University]. https://fenix.tecnico.ulisboa.pt/cursos/mma/dissertacao/1691203502343801

Omar, A., Delnaz, A., & Nik-Bakht, M. (2023). Comparative analysis of machine learning techniques for predicting water main failures in the City of Kitchener. *Journal of Infrastructure Intelligence and Resilience*, *2*(3). https://doi.org/10.1016/j.iintel.2023.100044

Paradkar, A. B. (2012). *An evaluation of failure modes for cast iron and ductile iron water pipes* [Master's thesis, University of Texas at Arlington]. https://rc.library.uta.edu/uta-ir/handle/10106/11660

Phan, H. C., Dhar, A. S., Hu, G., & Sadiq, R. (2019). Managing water main breaks in distribution networks—A risk-based decision making. *Reliability Engineering and System Safety*, *191*. https://doi.org/10.1016/j.ress.2019.106581

Plastic pipework. (n.d.). Wikipedia. Retrieved December 15, 2020, from https://en.wikipedia.org/wiki/Plastic_pipework

Rahbaralam, M., Modesto, D., Cardus, J., Abdollahi, A., & Cucchietti, F. (2007). *Predictive analytics for water asset management: Machine learning and survival analysis*. arxiv:2007.03744 [eess.SP]

Rahman, S., & Vanier, D. J. (2004). *An evaluation of condition assessment protocols for sewer management* (Issue B-5123.6). https://doi.org/10.4224/20377409

Rajani, B., & Kleiner, Y. (2012). Fatigue failure of large-diameter cast iron mains. *Water Distribution Systems Analysis 2010 - Proceedings of the 12th International Conference, WDSA 2010*. https://doi.org/10.1061/41203(425)104

Rajani, B., & Makar, J. (2000). A methodology to estimate remaining service life of grey cast iron water mains. *Canadian Journal of Civil Engineering*, *27*(6), 1259–1272. https://doi.org/10.1139/l00-073

Rajani, B., McDonald, S., & Félio, G. (1993). *Water mains break data on different pipe materials for 1992 and 1993 (A-7019.1 Final)*.

Ravichandran, T., Gavahi, K., Ponnambalam, K., Burtea, V., & Mousavi, J. S. (2021). Ensemble-based machine learning approach for improved leak detection in water mains. *Journal of Hydroinformatics*, *23*(2). https://doi.org/10.2166/HYDRO.2021.093

Rezaei, H., Ryan, B., & Stoianov, I. (2015). Pipe failure analysis and impact of dynamic hydraulic

conditions in water supply networks. *Procedia Engineering*, *119*(1), 253–262. https://doi.org/10.1016/j.proeng.2015.08.883

Robles-Velasco, A., Cortés, P., Muñuzuri, J., & De Baets, B. (2023). Prediction of pipe failures in water supply networks for longer time periods through multi-label classification. *Expert Systems with Applications*, *213*. https://doi.org/10.1016/j.eswa.2022.119050

Robles-Velasco, A., Cortés, P., Muñuzuri, J., & Onieva, L. (2020). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. *Reliability Engineering and System Safety*, *196*. https://doi.org/10.1016/j.ress.2019.106754

Romer, A. E., Bell, G. E. C., & Ellison, R. D. (2007). Failure of prestressed concrete cylinder pipe. *Pipelines 2007: Advances and Experiences with Trenchless Pipeline Projects - Proceedings of the ASCE International Conference on Pipeline Engineering and Construction*. https://doi.org/10.1061/40934(252)64

Royer, M. D. (2005). White paper on improvement of structural integrity monitoring for drinking water mains. *U.S. Environmental Protection Agency*.

Scheidegger, A., Leitão, J. P., & Scholten, L. (2015). Statistical failure models for water distribution pipes - A review from a unified perspective. In *Water Research* (Vol. 83). https://doi.org/10.1016/j.watres.2015.06.027

Scheidegger, A., Scholten, L., Maurer, M., & Reichert, P. (2013). Extension of pipe failure models to consider the absence of data from replaced pipes. *Water Research*, *47*(11). https://doi.org/10.1016/j.watres.2013.04.017

Selvadurai, A. P. S., & Shinde, S. B. (1993). Frost heave induced mechanics of buried pipelines. *Journal of Geotechnical Engineering*, *119*(12), 1929–1951. https://doi.org/10.1061/(ASCE)0733-9410(1993)119:12(1929)

Singh, A., & Adachi, S. (2013). Bathtub curves and pipe prioritization based on failure rate. *Built Environment Project and Asset Management*, *3*(1). https://doi.org/10.1108/BEPAM-11-2011-0027

Snider, B., & McBean, E. A. (2018). Improving time-to-failure predictions for water distribution systems using gradient boosting algorithm. *1st International WDSA / CCWI 2018 Joint Conference*.

Snider, B., & McBean, E. A. (2020a). Improving Urban Water Security through Pipe-Break Prediction Models: Machine Learning or Survival Analysis. *Journal of Environmental Engineering*, *146*(3). https://doi.org/10.1061/(asce)ee.1943-7870.0001657

Snider, B., & McBean, E. A. (2020b). Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions. *Urban Water Journal*, *17*(2), 163–176. https://doi.org/10.1080/1573062X.2020.1748664

Snider, B., & McBean, E. A. (2021). Combining Machine Learning and Survival Statistics to Predict Remaining Service Life of Watermains. *Journal of Infrastructure Systems*, *27*(3). https://doi.org/10.1061/(asce)is.1943-555x.0000629

Stone, S. L., Dzuray, E. J., Meisegeier, D., Dahlborg, a. S., Erickson, M., National Risk Management Research Laboratory (US), & Logistics Management Institute. (2002). Decision-support tools for predicting the performance of water distribution and wastewater collection systems. *Environmental Protection*, 1–101. http://www.epa.gov/ordntrnt/ORD/NRMRL/pubs/600r02029/600R02029.pdf

Sundahl, A. C. (1996). *Diagnosis of water pipe conditions* [Doctoral dissertation, Lund University]. ISSN 1101-9824.

Turkson, A. J., Ayiah-Mensah, F., & Nimoh, V. (2021). Handling censoring and censored data in survival analysis: A standalone systematic literature review. *International Journal of Mathematics and Mathematical Sciences*, *2021*. https://doi.org/10.1155/2021/9307475

Vairavamoorthy, K., & Ali, M. (2005). Pipe index vector: A method to improve genetic-algorithm-based pipe optimization. *Journal of Hydraulic Engineering*, *131*(12). https://doi.org/10.1061/(asce)0733-9429(2005)131:12(1117)

Vapnik, V. N. (1995). The nature of statistical learning theory. In *The Nature of Statistical Learning Theory*. https://doi.org/10.1007/978-1-4757-2440-0

Velasco Robles, A. (2022). *A machine learning approach to predict pipe failures in water distribution networks*. https://dialnet.unirioja.es/servlet/dctes?codigo=305976

Vishwakarma, A., & Sinha, S. K. (2020). Development of a consequence of failure model and risk matrix for water pipelines infrastructure systems. *Pipelines 2020: Utility Engineering, Surveying, and Multidisciplinary Topics - Proceedings of Sessions of the Pipelines 2020*

Conference. https://doi.org/10.1061/9780784483213.019

Wald, A. (1943). A method of estimating plane vulnerability based on damage of survivors. *Center for Naval Analyses*.

Walker, B. (2011). North America's Cinderella pipe story: A look at PVC pipes' climb to the top. *Journal of ASTM International*, *8*(7). https://doi.org/10.1520/JAI102839

Weeraddana, D., Liang, B., Li, Z., Wang, Y., Chen, F., Bonazzi, L., Phillips, D., & Saxena, N. (2018). Utilizing machine learning to prevent water main breaks by understanding pipeline failure drivers. *Ozwater*. http://arxiv.org/abs/2006.03385

Weeraddana, D., Liang, B., Li, Z., Wang, Y., Chen, F., Bonazzi, L., Phillips, D., & Saxena, N. (2019). Machine learning for water mains maintenance. *Water E-Journal*, *4*(3), 1–13. https://doi.org/10.21139/wej.2019.017

Wilson, D., Filion, Y., & Moore, I. (2017). State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. *Urban Water Journal*, *14*(2). https://doi.org/10.1080/1573062X.2015.1080848

Wilson, D., Moore, I., & Filion, Y. (2014). Development of a methodology to predict the failure of large-diameter cast iron water mains. *World Environmental and Water Resources Congress 2014: Water Without Borders - Proceedings of the 2014 World Environmental and Water Resources Congress*. https://doi.org/10.1061/9780784413548.048

Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W., & Tscheikner-Gratl, F. (2018). Pipe failure modelling for water distribution networks using boosted decision trees. *Structure and Infrastructure Engineering*, *14*(10). https://doi.org/10.1080/15732479.2018.1443145

Wu, Y., Kang, C., Nojumi, M. M., Bayat, A., & Bontus, G. (2021). Current water main rehabilitation practice using trenchless technology. *Water Practice and Technology*, *16*(3). https://doi.org/10.2166/wpt.2021.026

Xu, H., & Sinha, S. K. (2020). Applying survival analysis to pipeline data: Gaps and challenges. *Pipelines 2020: Utility Engineering, Surveying, and Multidisciplinary Topics - Proceedings of Sessions of the Pipelines 2020 Conference*. https://doi.org/10.1061/9780784483213.017

Xu, H., & Sinha, S. K. (2021). Modeling pipe break data using survival analysis with machine

learning imputation methods. *Journal of Performance of Constructed Facilities*, *35*(5). https://doi.org/10.1061/(asce)cf.1943-5509.0001649

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/775047.775151

Zakikhani, K., Nasiri, F., & Zayed, T. (2021). A failure prediction model for corrosion in gas transmission pipelines. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, *235*(3). https://doi.org/10.1177/1748006X20976802

Zarghamee, M. S., Ojdrovic, R. P., & Nardini, P. D. (2011). Prestressed concrete cylinder pipe condition assessment what works, what doesn't, what's next. *Pipelines 2011: A Sound Conduit for Sharing Solutions - Proceedings of the Pipelines 2011 Conference*. https://doi.org/10.1061/41187(420)18

Zhang, B., Guo, T., Zhang, L., Lin, P., Wang, Y., Zhou, J., & Chen, F. (2018). *Water pipe failure prediction: A machine learning approach enhanced by domain knowledge*. https://doi.org/10.1007/978-3-319-90403-0_18

Zhou, X., Tang, Z., Xu, W., Meng, F., Chu, X., Xin, K., & Fu, G. (2019). Deep learning identifies accurate burst locations in water distribution networks. *Water Research*, *166*, 115058. https://doi.org/10.1016/j.watres.2019.115058

Appendices

Appendix A
Detailed Data Tables

A.1. Pipe Material Detail

Cast Iron (CI)

Description	Older, thick-walled ferrous pipes. While susceptible to corrosion, the sheer volume of metal used offered these pipes a long service life.
Degradation	1) Internal and external corrosion, both general and pitting corrosion. 2) Cracking of the pipe wall (brittle failure). 3) Joint failure, particularly where joints were sealed using a material with a different coefficient of thermal expansion, such as leadite.
Usage	Common for small, medium, and large diameters prior to 1960 (Paradkar, 2012).
Assessment	Measuring the loss of pipe wall thickness via a range of techniques. Leak detection often used to detect joint failure.
Rehabilitation	Cleaning and lining, often with cement mortar, can extend the life of these pipes.

Ductile Iron (DI)

Description	Newer, thinner walled ferrous pipes. Often internally lined and externally coated to provide protection against corrosion.
Degradation	1) Corrosion pitting where external coating is damaged. 2) General corrosion where no corrosion protection applied.
Usage	Common for small and medium diameters from 1960 to 2000 (Paradkar, 2012).
Assessment	Leak detection for pinhole leaks due to corrosion pitting. Acoustic wall thickness testing for general corrosion. Inline electromagnetic techniques are effective but are rarely used due to high ratio of inspection cost to replacement cost.
Rehabilitation	Cathodic protection via sacrificial anodes.

Steel (ST)

Description	High strength pipe, more expensive than iron. Generally internally lined and externally coated to provide protection against corrosion.
Degradation	Corrosion pitting where internal lining or external coating is damaged.
Usage	Medium and large diameters, particularly for high pressures (Joshi, 2012).
Assessment	Leak detection for pinhole leaks due to corrosion pitting. Inline near field eddy current and magnetic flux leakage are sometimes used for large diameter pipes that are highly critical or have a high cost of replacement.
Rehabilitation	1) Cathodic protection via impressed current.

	2) Replacement or carbon fiber reinforcement of individual degraded sections.

Asbestos Cement (AC)

Description	Cement mixed with asbestos fibers to provide tensile strength.
Degradation	1) Calcium leaches from the material, weakening the pipe wall. 2) Cracking, particular in weakened portions (brittle failure).
Usage	Common for small and medium diameters from 1940 to 1970 (Hu et al., 2008).
Assessment	1) Acoustic wall thickness testing to measure average loss of structural material. 2) Phenolphthalein dye testing to spot check loss of structural material.
Rehabilitation	Uncommon; degraded pipe is usually replaced.

Polyvinyl Chloride (PVC)

Description	Newer plastic pipes using a variety of materials.
Degradation	These pipe types are relatively new and have low rates of leakage and failures. Aside from the first few small cohorts made from substandard material, little research has been published on the actual degradation of plastic pipes.
Usage	Increasingly common for small diameters from 1950 onwards (*Plastic Pipework*, n.d.).
Assessment	No common and effective method is available.
Rehabilitation	Uncommon. These pipes are generally relatively new.

Polyethylene (PE), including High Density Polyethylene (HDPE)

Description	Newer plastic pipes using a variety of materials.
Degradation	These pipe types are relatively new and have low rates of leakage and failures. Aside from the first few small cohorts made from substandard material, little research has been published on the actual degradation of plastic pipes.
Usage	Increasingly common for small diameters from 1950 onwards (*Plastic Pipework*, n.d.).
Assessment	No common and effective method is available.
Rehabilitation	Uncommon. These pipes are generally relatively new.

Concrete (CONC), including Prestressed Concrete Cylinder Pipe (PCCP)

Description	Composite pipe made from a concrete core wrapped with prestressing wires to keep it under compression. A steel cylinder is often used inside the wires to act as a watertight membrane. Cement mortar is applied for corrosion protection on the inside and outside.

Degradation	Prestressing wires may suffer from corrosion or hydrogen embrittlement and then snap. Once sufficient prestress is lost, the concrete core can lose its compression, allowing the pipe to rupture catastrophically.
Usage	Common for large diameters 1950 onwards (Manda, 2012).
Assessment	Testing for breaks in the prestressing wire via transformer coupling or acoustic monitoring.
Rehabilitation	Replacement or carbon fiber reinforcement of individual degraded sections.

Copper (CU)

Description	Flexible metal pipe made almost exclusively in small diameters.
Degradation	External corrosion; internal mineral buildup reducing capacity.
Usage	Widely used for small diameters (Groover, 2019).
Assessment	Direct visual assessment.
Rehabilitation	Uncommon. These pipes are generally replaced.

A.2. Pipe Condition Assessment Methods

Pit Depth Measurement

Description	A pit gauge is used to manually measure the depths of corrosion pits. The surface is often prepared by sandblasting to remove rust and graphitized material.
Applicability	Quantify loss of wall thickness from corrosion in ferrous pipes. Primarily used on cast iron pipe and to a lesser extent ductile iron and steel.
Usage	Localized measurements, either from coupons or external in-situ.
Raw Data	Maximum or average pit depth(s) within given surface areas.
Analysis	Fit a statistical curve of pit depth populations and extrapolate.
Analyzed Data	Shape & location parameters of statistical curve or extrapolated pit statistics for full pipe segments.

Ultrasonic Testing

Description	An ultrasonic pulse is applied to the pipe wall. This reflects off the opposite surface, and the propagation time is used to calculate the wall thickness.
Applicability	Quantify loss of wall thickness from corrosion in ferrous pipes. Sometimes used in conjunction with pit depth measurements to estimate original wall thickness.
Usage	1) Localized measurements, either from coupons or external in-situ. 2) Inline testing.
Raw Data	Ultrasonic pulse propagation times.
Analysis	Calculate wall thickness via speed of sound in material.

Analyzed Data	Wall thickness at measurement points. Note: Presence of graphitized material causes multiple reflections, making analysis more difficult and possibly leading to false readings.

Near Field Eddy Current

Description	An electromagnetic testing method where a periodic signal is applied to the pipe surface. Localized eddy currents are measured beneath an array of small sensors, allowing determination of average all thickness beneath each. Known as Broadband Electromagnetic when multiple frequencies used concurrently.
Applicability	Quantify loss of wall thickness from corrosion in ferrous pipes. Less sensitive than magnetic flux leakage but lower cost and easier to apply, as it permits separation between sensor and pipe wall.
Usage	1) Localized measurements, usually external in-situ. 2) Inline testing (less common).
Raw Data	Phase and amplitude (or in-phase and quadrature) per frequency for each sensor.
Analysis	Relate the phase and amplitude to various thicknesses of the material (requires calibration to the specific material being tested).
Analyzed Data	Average wall thickness at each sensor for each measurement.

Magnetic Flux Leakage

Description	Magnets are placed close together in contact with the pipe wall, saturating the area between them with magnetic flux. Corrosion pits will cause leakage of this flux, which is detected by an array of Hall sensors close to the pipe wall.
Applicability	Quantify loss of wall thickness from corrosion in ferrous pipes. More sensitive than eddy current techniques but also more costly, as it requires close contact between sensor and pipe wall.
Usage	1) Localized measurements, usually external in-situ. 2) Inline testing (less common).
Raw Data	Unidirectional flux from each Hall sensor, sometimes supplemented by eddy current sensor to help determine if the flux originates from the inside or outside surface of the pipe wall.
Analysis	Calculate pit size based the measured flux.
Analyzed Data	Pit size per measurement. Note: second level analysis may be performed as per pit depth measurements.

Remote Field Eddy Current

Description	A periodic electromagnetic signal is applied to a pipe wall. Eddy currents propagate further than the direct signal, allowing these to be measured by multiple receiver sensors spaced around the inside of the pipe in the “remote field” distance (typically 10 to 15 pipe diameters away).

Applicability	Quantify loss of wall thickness from corrosion in ferrous pipes.
Usage	Inline testing.
Raw Data	Phase and amplitude of the received signal at each sensor.
Analysis	Calculate corrosion pit depths from variations in signal amplitude and phase.
Analyzed Data	Corrosion pit depths and locations.

Transformer Coupling

Description	A coil of wire (the "exciter") is used to induce a signal in the spiral-wound prestressing wire. The resulting signal is then measured by a second coil (the "detector"). Discontinuities in the prestressing wire cause changes in the signal.
Applicability	Prestressed concrete cylinder pipe (PCCP) to detect and quantify breaks in the spiral-wound prestressing wire.
Usage	Inline testing.
Raw Data	Phase and amplitude of the received signal.
Analysis	Calculate number of broken wires from phase and amplitude of received signal.
Analyzed Data	Number of broken wires.

Wire Break Monitoring

Description	Acoustic sensors are placed along the length of a pipe. When prestressing wires in the pipe break, this sound is captured as an indication of degradation.
Applicability	Prestressed concrete cylinder pipe (PCCP) to detect and quantify ongoing breakage of the spiral-wound prestressing wire.
Usage	External and internal hydrophones or internal fiber optic sensor.
Raw Data	Acoustic data.
Analysis	Detect wire break events in the acoustic data and quantify them.
Analyzed Data	Number of wires that broke during the monitoring period. Note: This measures the rate of degradation not the current state.

Acoustic Wall Thickness Testing

Description	Acoustic noise or a pressure wave is induced in the fluid inside a pipe. The propagation time for this disturbance is measured between two sensors located along the pipe. That, in turn, is used to calculate the average wall thickness over the interval between the two sensors.
Applicability	Primarily asbestos cement and cast iron pipe to measure the average remaining thickness of structural material in the pipe wall.
Usage	External testing, measuring average pipe condition over an interval (typically 100m to 1,000m).
Raw Data	Transit time and distance between a pair of sensors.

Analysis	Calculate average pipe wall thickness using relationship between speed of sound in a fluid-bearing pipe and the thickness of the pipe wall.
Analyzed Data	Average thickness of structural pipe wall material between a pair of sensors.

Phenolphthalein Dye Testing

Description	A cross-section of asbestos cement pipe is cut and smoothed; then it has a phenolphthalein dye applied. This dye turns pink in the presence of structurally sound material, allowing estimation of the remaining pipe wall strength.
Applicability	Quantify loss of effective wall thickness due to calcium leaching in asbestos cement pipes.
Usage	Local testing from pipe coupons.
Raw Data	Photographs of the pipe cross-section surface after application of dye.
Analysis	Measurement of loss of structural material from the internal and external diameter.
Analyzed Data	Average or maximum thickness loss from inner and outer diameter.

Leak Detection

Description	Acoustic sensors are placed at locations along or inside the pipe. The sound of water escaping is detected, indicating the presence and location of a leak.
Applicability	All pipe materials. Particularly applicable to ductile materials, such as ductile iron, steel, and PVC, which can support leaks for long periods before failing completely.
Usage	External sensors, or inline testing for greater sensitivity to small leaks.
Raw Data	Acoustic data.
Analysis	Sound amplitude, spectrum, and cross-correlation to detect & localize leaks.
Analyzed Data	Leak presence, sometimes location and size.

Inline Video (Closed Circuit Television)

Description	A closed-circuit television camera passed through the pipe to detect visible signs of degradation.
Applicability	Non-pressurized pipes running partially full (very rare in water networks).
Usage	Inline testing.
Raw Data	Video.
Analysis	Visual observation of defects such as cracks, obstructions, scouring, etc.
Analyzed Data	Defect types, severity, and locations per a standard defect coding system, often the WRC / NASSCO system (Rahman & Vanier, 2004).

A.3. Data Set Summary Statistics

The following tables in this appendix provide summary statistics of the data set. These are presented for the Pipes (**Error! Reference source not found.**), for the Breaks (**Error! Reference source not found.**), and for the Segment-Year combinations (**Error! Reference source not found.**). Measures are provided per utility contributor to provide a sense of the completeness of each contribution. One aggregate measure across all data sets is provided for each data column, with the aggregation selected to provide insight into the column. Note that for each column with a _count suffix, this measure provides the count of non-blank entries for that column. The sum of these entries describes how many of the entries include this feature.

Table 37: Descriptive statistics for the Pipes table, aggregated by participating utility.

Column	aw	ham	peel	sin	tor	wnet	Aggregation	
Count	77,256	36,769	57,376	208,842	50,401	151,330	Sum:	581,974
pipeid_distinct	77,256	36,769	57,376	208,842	50,400	151,330	Sum:	581,974
dia_min	19	1	10	15	0	1	Min:	0
dia_max	1,800	2,250	2,550	2,200	4,500	1,500	Max:	4,500
dia_count	77,256	36,769	57,376	208,842	50,383	150,290	Sum:	580,916
dia_avg	231	229	251	229	244	376	Mean:	260
dia_stddev	163	150	184	217	210	469	Mean:	232
mat_count	77,256	36,769	57,376	208,842	49,340	150,250	Sum:	579,833
mat_distinct	10	20	26	14	21	18	Sum:	109
len_min	0.01	0.00	0.00	0.22	0.10	0.00	Min:	0
len_max	2,136	2,320	5,478	7,034	3,855	4,855	Max:	7,034
len_avg (m)	66	57	95	41	131	25	Mean:	69
len_sum (km)	5,128	2,112	5,454	8,513	6,604	3,723	Sum:	31,535
len_stddev	83	86	149	98	151	52	Mean:	103
lining_count	0	3,075	0	167,617	17,974	106,208	Sum:	294,874
lining_distinct	0	3	0	9	2	8	Sum:	22
cor_protect_count	0	0	0	0	8,131	115,604	Sum:	123,735
cor_protect_distinct	0	0	0	0	1	7	Sum:	8
joints_count	0	0	0	0	12,379	115,403	Sum:	127,782
joints_distinct	0	0	0	0	11	18	Sum:	29
main_type_count	0	0	0	208,842	50,401	151,303	Sum:	410,546
main_type_distinct	0	0	0	3	3	12	Sum:	18
inst_year_min	1800	1758	1855	1900	1858	1853	Min:	1758
inst_year_max	2022	2020	2021	2022	2015	2021	Max:	2022
inst_year_avg	1970	1977	1994	1981	1954	1979	Mean:	1976

rem_year_min				1900	1913	209	Min:	209
rem_year_max				2021	2014	2021	Max:	2,021
rem_year_count	0	0	0	22,854	4,943	8,364	Sum:	36,161
rehab_year_min		1950		2006	2	1978	Min:	2
rehab_year_max		2019		2021	9991	2017	Max:	9,991
rehab_year_count	0	2,937	0	48	20,313	1,163	Sum:	24,461
active_count	0	0	57,376	208,842	50,401	151,330	Sum:	467,949
active_sum			48,210	156,284	44,623	140,039	Sum:	389,156
loc1_count	77,256	0	0	183,042	50,393	151,293	Sum:	461,984
loc2_count	0	0	0	0	0	151,330	Sum:	151,330
owner_count	0	0	57,376	0	50,401	151,316	Sum:	259,093
owner_distinct	0	0	7	0	5	2	Sum:	14
water_type_count	0	0	0	208,842	0	151,316	Sum:	360,158
water_type_distinct	0	0	0	3	0	3	Sum:	6
note1_count	0	0	0	208,733	8,894	29,267	Sum:	246,894
roughness_count	0	0	0	191,392	0	151,330	Sum:	342,722
roughness_distinct	0	0	0	22	0	13	Sum:	35
pressure_class_count	0	0	0	0	0	116,729	Sum:	116,729
pressure_class_distinct	0	0	0	0	0	13	Sum:	13
manufacturer_count	0	0	0	0	0	37,076	Sum:	37,076
manufacturer_distinct	0	0	0	0	0	16	Sum:	16
deadend_count	0	0	0	0	5,185	0	Sum:	5,185
deadend_distinct	0	0	0	0	2	0	Sum:	2
soil_type_count	0	27,825	57,272	0	0	0	Sum:	85,097
soil_type_distinct	0	14	8	0	0	0	Sum:	22
soil_depth_count	0	36,769	0	0	0	0	Sum:	36,769

The Pipe table values contain many obvious data quality issues. An example is the "rem_year_min" (earliest year in which a pipe was removed from service) for wnet in the Pipes table. This shows the year 209, and is likely a data entry error where 2009 was intended. A similar issue presents in the "rehab_year_max" (latest year in which a pipe was rehabilitated) for tor, which shows 9991 (likely 1991). A deliberate choice was made to leave these entries as they are when they do not interfere with the data processing pipelines. This maintains an accurate representation of the expected outcomes of attempting to apply this model to a new utility's data.

Table 38: Descriptive statistics for the Breaks table, aggregated by participating utility.

Column	aw	ham	peel	sin	tor	wnet	Aggregation	
count	4,362	7,345	7,643	10,330	76,027	48,275	Sum:	153,982
brid_distinct	4,362	7,344	7,643	10,330	76,027	48,162	Sum:	153,868
pipeid_distinct	3,314	3,683	3,263	3,619	17,011	19,777	Sum:	50,667
date_count	4,362	7,345	7,643	10,330	76,027	48,275	Sum:	153,982
br_year_min	2010	1982	1975	2010	1963	1999	Min:	1,963
br_year_max	2022	2019	2020	2021	2014	2019	Max:	2,022
br_year_distinct	14	39	48	13	56	22	Mean:	32
loc1_count	4,362	0	0	4,610	69,961	47,909	Sum:	126,842
loc2_count	0	0	0	0	70,884	47,900	Sum:	118,784
event_class_count	0	0	0	0	0	48,275	Sum:	48,275
event_class_distinct	0	0	0	0	0	1	Sum:	1
cause_count	0	0	0	10,330	61,369	0	Sum:	71,699
cause_distinct	0	0	0	3	198	0	Sum:	201
note1_count	0	7,345	0	0	12,392	48,056	Sum:	67,793
isbr_count	4,362	0	0	0	0	48,275	Sum:	52,637
break_type_count	2,089	0	0	0	61,859	0	Sum:	63,948
break_type_distinct	4	0	0	0	8	0	Sum:	12
temp_count	0	0	0	0	34,574	0	Sum:	34,574
fix_count	0	0	0	0	24,300	0	Sum:	24,300
fix_distinct	0	0	0	0	7,117	0	Sum:	7,117
soil_depth_count	0	0	0	0	55,133	0	Sum:	55,133
soil_type_count	0	6,949	0	0	8,323	0	Sum:	15,272
soil_type_distinct	0	13	0	0	11	0	Sum:	24

The segment-years table is generated after the data integration steps described in Chapter 3.3, and as such shows statistics from after certain minimal data cleansing activities were undertaken. As an example, manual examination of the tor dataset showed that recording of pipe breaks begin as an inconsistent ad-hoc activity, with some years having zero pipe breaks recorded. The activity appears to have become consistent around 1963 and was likely added to the city's standard procedures around this time. To avoid generating Segment-Year records from the period of 1928 to 1962 that incorrectly show very few breaks, all data from prior to 1963 was dropped. Similarly, each dataset included break records spanning only a portion of the final year. To avoid creating artificially low break rates for the final year of records for each dataset (and the pipe segments and cohorts contained therein), this final partial year was dropped.

Table 39: Descriptive statistics for the Segment-Years table, aggregated by participating utility.

Column	aw	ham	peel	sin	tor	wnet	Aggregation	
count	841,142	1,026,773	1,417,439	1,875,295	2,122,983	2,762,929	Sum:	10,046,561
year_min	2010	1982	1975	2010	1963	1999	Min:	1963
year_max	2022	2019	2020	2021	2014	2019	Max:	2022
pipeid_distinct	72,019	36,635	57,316	175,954	50,305	147,147	Sum:	539,376
breaks_this_year_sum	3,000	7,228	7,104	4,132	68,277	37,027	Sum:	126,768
had_break_sum	2,845	6,697	6,198	3,830	60,460	30,825	Sum:	110,855
dia_avg	240	227	252	242	239	413	Mean:	269
len_avg (m)	70	60	103	38	135	24	Mean:	72
len_sum (km)	59,200	61,697	146,245	71,907	286,086	67,596	Sum:	692,732
age_avg	39	37	19	23	46	34	Mean:	33
Breaks_per_100_km_per_year	5.07	11.72	4.86	5.75	23.87	54.78	Mean:	17.67
Breaks_per_1000_seg_years	3.57	7.04	5.01	2.20	32.16	13.40	Mean:	10.56

Two measures of the break rate are provided for each dataset: breaks per 100 km per year and breaks per 1,000 segments per year. Breaks per 100 km per year is a common metric in the industry. Breaks per 1,000 segments per year was selected to provide a scaling that is on a similar order of magnitude, given the average segment lengths in the data set. These break rates show considerable variation from one utility to another.