

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/340621434>

# Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions

Article in *Urban Water Journal* · April 2020

DOI: 10.1080/1573062X.2020.1748664

CITATIONS

0

2 authors:



Brett Snider

University of Guelph

7 PUBLICATIONS 7 CITATIONS

[SEE PROFILE](#)

READS

17



Edward McBean

University of Guelph

386 PUBLICATIONS 4,107 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Framework and Methodology for Improved Indigenous-led Decision-Making on Water and Wastewater Management and Design [View project](#)



Energy Performance of Water Distribution Systems [View project](#)



## Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions

Brett Snider & Edward A. McBean

To cite this article: Brett Snider & Edward A. McBean (2020): Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions, Urban Water Journal

To link to this article: <https://doi.org/10.1080/1573062X.2020.1748664>



Published online: 14 Apr 2020.



Submit your article to this journal [↗](#)



View related articles [↗](#)




View Crossmark data [↗](#)

RESEARCH ARTICLE



# Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions

Brett Snider  and Edward A. McBean

School of Engineering, University of Guelph, Guelph, Ontario, Canada

## ABSTRACT

Many water utilities are facing a crisis of aging infrastructure. Aging pipes are deteriorating, and pipe breaks are increasing. A variety of pipe break prediction models have been developed to identifying which pipes are most likely to break next, in order to assist utilities in prioritizing pipe replacement. This paper investigates the role of data in pipe break prediction model accuracy. A gradient boosting decision tree machine learning model, a Weibull proportional hazard probabilistic model and two ranking models (based on 'age of pipe' and 'previous-break') were calibrated using a various number of pipes, years of break records and input variables. The results indicate how the different model types are impacted by data limitations. Overall, this study finds the Age-based approach to be inaccurate, and the XGBoost machine learning model demonstrates superior predictive capability when the training dataset contains more than 5 years of break records and 2,000 or more pipes.

## ARTICLE HISTORY

Received 25 September 2019  
Accepted 25 March 2020

## KEYWORDS

Water main failure; statistical models; asset management; machine learning; data mining; infrastructure

## 1. Introduction

Watermain failure is a major concern for every water utility. Watermain breaks, also referred to in this paper as pipe breaks, disrupt customer service, result in water and revenue loss, and create the potential for contaminants to enter the water distribution system. The total cost of water loss due to watermain or pipe breaks is estimated to be 3.8 USD billion per year in North America (EPA 2010; Renzetti and Dupont 2013). Moreover, this value increases dramatically when including indirect costs, such as interruption to service, and health impacts (Renzetti and Dupont 2013). At the same time, pipe breaks and their associated costs are expected to increase over the next few decades as many of North America's watermains approach their expected end-of-service life (AWWA 2012). In light of these challenges, the development of strategies to reduce pipe breakage is crucial, particularly as the world faces numerous crises related to water security and sustainability.

Prediction models can help utilities reduce future breaks by identifying which pipes are most likely to break, and when. Utilities are able to use these predictions to develop more effective asset management plans and replace pipes before major breaks occur. Researchers have developed a variety of watermain break prediction models that rely on historical data such as pipe break records, pipe attributes, and hydraulic models (e.g. Kleiner and Rajani 2001; Wilson, Filion, and Moore 2015). When these models are calibrated using large historic databases, they are able to predict which pipes are most likely to break next, and thus assist utilities in designing effective pipe repair/replacement strategies. However, how these models perform when faced with limited break records has not yet been explored. Without this information, researchers and utilities are

unsure as to which pipe break prediction model is most appropriate for their specific data availability.

This paper explores the role historical databases play in building accurate pipe break prediction models. Specifically, the data requirements for four break models are studied; two simplistic ranking models, a machine learning model, and a probabilistic model. The models are calibrated using varying numbers of pipes, input attributes and years of break records. The results of these comparisons outline the role data plays in building accurate pipe break prediction models and provides guidance to utilities in selecting the appropriate model based on the availability of data.

This research is needed now, more than ever, given the burgeoning infrastructure crisis affecting utilities throughout North America. In North America 28% of the pipes that are in service today are over 50 years of age (Folkman 2018) and are approaching their expected end-of-service. The impacts of this deteriorating infrastructure are hard to ignore, with water main breaks increasing by 27% in the last six years (Folkman 2018). To combat this, major pipe replacement is needed. However, pipe replacement comes with a heavy price tag, with an estimated 500 USD billion being required to replace the aging infrastructure over the next 25 years (AWWA 2012).

Pipe break prediction models can assist utilities in identifying which pipes need to be replaced in order to reduce future breaks and avoid replacing pipes that are still in good condition. The cost savings associated with adopting more accurate pipe break prediction models can be substantial. For example, one case study suggests that by adopting a machine learning pipe break prediction model, instead of replacing pipes based solely on their age, the utility is able to reduce the length of pipe replaced by four miles while maintaining the same number of breaks (Hatler 2019). The cost savings associated with not

replacing these four miles of pipe is estimated to avoid 4 USD million in expenses (Hatler 2019).

Hatler's (2019) case study demonstrates the importance of building an effective pipe break prediction model, but the relationship between the data used to calibrate the model and the model's accuracy remains undefined. Without this information, utilities are left to 'guess' which model is most appropriate, given the data availability for their system. This paper addresses this issue by comparing commonly used pipe break prediction models calibrated with various amounts of data, in order to identify which model is most accurate based on a utility's unique data availability.

## 2. Pipe break prediction models

Before investigating the role data plays in pipe break prediction models, it is important to understand the various pipe break models that have been developed. Pipe break prediction models have been well described in various literature reviews (e.g. Kleiner and Rajani 2001; Rajani and Kleiner 2001; St. Clair and Sinha 2012; Nishiyama and Filion 2014; Scheidegger, Leitão, and Scholten 2015; Dawood et al. 2019). Each of these reviews have different classification methods to describe the various models. The classification of prediction models adopted in this paper is outlined in Figure 1.

### 2.1. Simplistic models

Typically neglected from literature reviews, simplistic models are often the most common approach adopted by utilities to predict pipe breakage and prioritize pipe replacement (Underhill and Elliott 2012). This paper considers simplistic models to only rank the likelihood of pipe breakage and to

not require calibration. The most common simplistic models rank pipes' likelihood of breakage based on the age of the watermain or the number of previous breaks. That is, the oldest pipes or the pipes with the largest number of previously recorded breaks are considered most likely to break next (Kirmeyer, Richards, and Smith 1994).

These simplistic models often rely on a single factor to rank the likelihood of pipe breakage, whereas an array of factors influence pipe breaks, including: soil properties, pipe material, diameter of pipe, water pressure, traffic loads, frost depths, etc. Therefore, pipe break prediction models that rely on a single factor such as age or number of previous breaks, may not be the best choice when trying to prioritize pipe replacement and identify which pipe is most likely to break next.

### 2.2. Physical models

Physical models attempt to describe the mechanisms that contribute to breaks by analysing the stressors acting on a pipe and comparing it to the estimated remaining strength of the pipe. These models often require specific, infield measurements to determine the stressors and remaining strength of the pipe. Measuring and calculating these values can be very time-consuming and expensive. For these reasons, physical pipe break prediction models are typically only used for critical watermains (Wilson, Filion, and Moore 2015). For an overview of physical models developed see Rajani and Kleiner (2001).

### 2.3. Statistical models

Statistical pipe break models apply statistical techniques to historical break data to identify patterns between input parameters and pipe breaks. Compared to physical models,

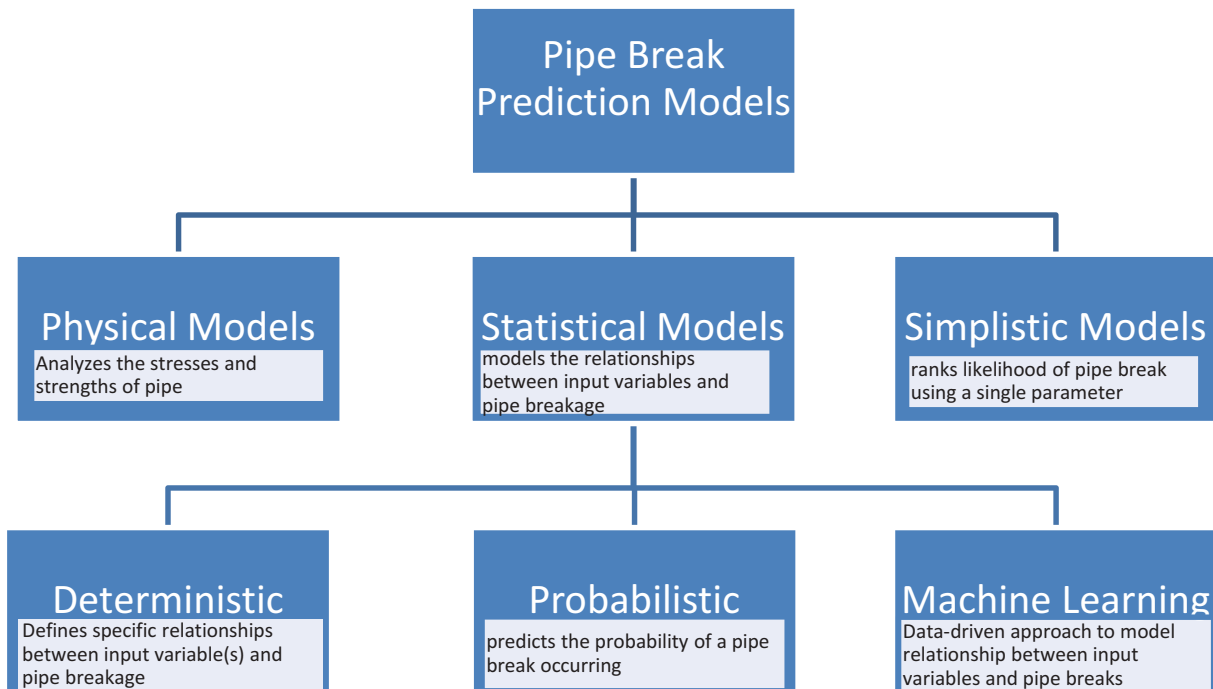


Figure 1. A taxonomy of pipe break prediction models.

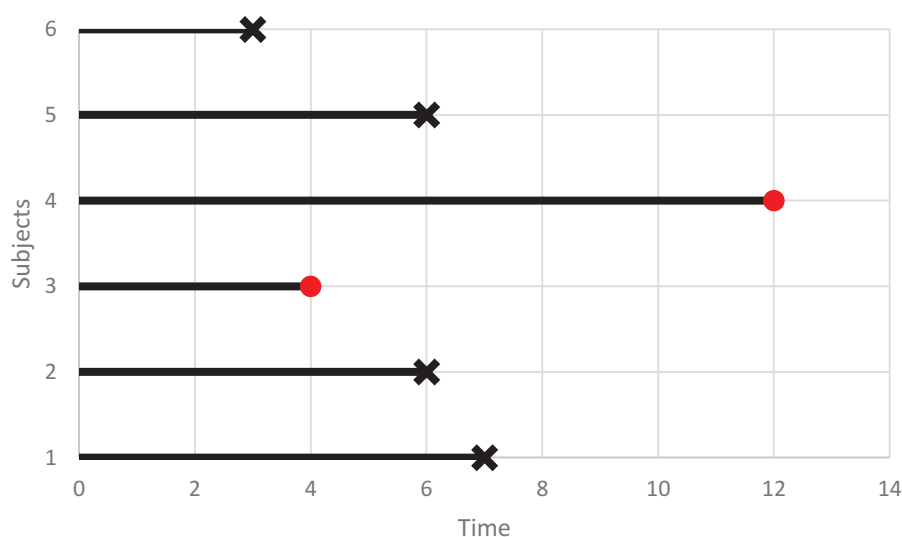


Figure 2. Censored data.

statistical models are less expensive and less time-consuming, and are thus the preferred model for the large numbers of distribution pipes within a water distribution system (Kleiner and Rajani 2001; Wilson, Fillion, and Moore 2015). Statistical pipe break models have been classified in numerous ways throughout the literature.

Often debated is whether to classify machine learning models as a type of statistical model or in a separate data-driven category. Nishiyama and Fillion (2014) argue ‘soft-computing’ models, which include machine learning models, should be grouped under statistical models, and Scheidegger, Leitão, and Scholten (2015) argue ‘artificial intelligence models such as ANN are also statistical models as they can be considered as highly flexible non-linear regression models’. Therefore, this paper employs the classification system adopted by these papers wherein machine learning/data-driven models are considered a subset of statistical models. Other statistical model classifications include deterministic and probabilistic pipe break prediction models.

### 2.3.1. Deterministic models

Deterministic models were the first statistical models developed to predict watermain pipe breaks. These models calculate input parameter coefficients by relying on regression techniques such as least-squares or maximum-likelihood estimation. The outputs from these models are specific break rates or time-to-break prediction and give very little indication regarding the probability of the outcome. Deterministic models developed for pipe break prediction include multivariate exponential regression model (Kleiner and Rajani 2000, 2002; Yamijala, Guikema, and Brumbelow 2009), Poisson-generalized linear regression model (Boxall et al. 2007; Asnaashari et al. 2009; Yamijala, Guikema, and Brumbelow 2009), multivariate linear regression model (Asnaashari et al. 2009; Wang, Zayed, and Moselhi 2009; Yamijala, Guikema, and Brumbelow 2009), and logistic generalized linear regression model (Yamijala, Guikema, and Brumbelow 2009).

Most of these models assume uniform breaks within watermain groups, and therefore require careful group selection.

Because of this limitation, and the inability of deterministic models to handle unreliable data, St. Clair and Sinha (2012) advise against adopting deterministic models for individual pipe break prediction.

### 2.3.2. Probabilistic models

Probabilistic models predict the probability of a pipe break occurring, based on extensive historical data. These models are designed to handle the randomness expected within pipe break databases. Probabilistic models include Yule process, Bayesian, and survival analysis models.

Of note and popular, are survival analysis models, which are designed to handle censored data inherent in pipe break databases. Censoring occurs when the event of interest is not observed within the dataset due to limited observation periods. For this reason, survival models are often used in medical trials where patient observation time is limited (e.g. Wiesner et al. 1989; Van Buuren, Boshuizen, and Knook 1999).

Censoring is also common in pipe break databases, where many of the pipes have yet to break and hence the event of interest, or break, is not yet observed. Figure 2 highlights an example of the censoring common in pipe break databases.

In Figure 2, Subjects 1 through 6 represent individual pipes. Subjects 1, 2, 5 and 6 are not considered censored since the subjects, or pipes, were observed from the beginning of the study (Time 0) to time of break (i.e. event of interest – marked ‘X’). Subjects 3 and 4 are considered ‘right censored’ since the break event did not occur within the study period. Survival analysis models incorporate such censored data into the model’s calibration, gleaned useful information from the censored event – particularly, that the actual event (i.e. pipe break) must occur after the censored event.

A large array of pipe break prediction models have been developed using survival analysis (Le Gat and Eisenbeis 2000; Vanrenterghem-Raven et al. 2004; Alvisi and Franchini 2010; Clark et al. 2010; Debón et al. 2010; Fuchs-Hanusch et al. 2012; Kimutai et al. 2015; Kleiner and Rajani 2012; Park et al. 2007, 2008, 2011). These models indicate improved pipe break prediction accuracy but rely on historical databases that are not

available for many utilities. Although the historical break records used to calibrate these models vary, from 6 years (Debón et al. 2010) to over 30 years (Kleiner and Rajani 2012), the specific relationship between historical databases and pipe break prediction accuracy has yet to be defined.

### 2.3.3. Machine learning models

The last several decades have witnessed significant advancements in machine learning algorithms and computational power. Unlike probabilistic and deterministic models which explicitly define covariate relationships, machine learning models adopt a data-driven approach to identify complex relationships between input and output variables. Due to this advantage, numerous machine learning models have been developed to predict future pipe breaks. These models include Artificial Neural Networks (Achim, Ghotb, and McManus 2007; Ahn et al. 2005; Asnaashari et al. 2013; Jafar, Shahrour, and Juran 2010; Tabesh et al. 2009), ensemble machine learning algorithms such as Random Forest and XGBoost (Winkler et al. 2018; Kumar et al. 2018; Snider and McBean 2018), genetic programming models (Xu, Chen, and Li 2011; Xu et al. 2011), and evolutionary polynomial regression (Berardi et al. 2008; Savic, Giustolisi, and Laucelli 2009; Xu et al. 2011; Laucelli et al. 2014) to name only a few.

Machine learning models have been shown to improve pipe break prediction accuracy (Achim, Ghotb, and McManus 2007). However, these models have difficulty including censored events. Therefore, the majority of machine learning time-to-break models simply remove censored pipe information from the dataset. This suggests that machine learning models may require larger historical databases in order to ensure that an adequate number of uncensored events are included in the training datasets. An investigation into how the number of pipes, pipe attributes and years of break records included in the training dataset impacts the various models' accuracy, has yet to be conducted.

## 3. Model selection

As detailed in the previous section, numerous pipe break prediction models have been developed, making it impractical to compare all models herein. Hence, this paper focuses on studying the impacts of limited data on four frequently used models that have been adopted by utilities to predict pipe breaks for water distribution mains. The models were selected to represent the different model classifications; simplistic, probabilistic, and machine learning models. No physical or deterministic models are demonstrated since these models are often impractical for individual water main pipe break prediction (St. Clair and Sinha 2012).

Two simplistic ranking models were selected, Age-based and Previous-break due to their wide-spread popularity amongst utilities with limited data (Kirmeyer, Richards, and Smith 1994). The Weibull proportional hazard model (Weibull PH) was chosen to represent probabilistic models since it has been shown to have superior accuracy over other probabilistic models (Kimutai et al. 2015). Lastly, the extreme gradient boosting (XGBoost) model was chosen to represent machine learning models since it has been shown to outperform other machine

learning models such as artificial neural networks (ANN) and random forests in predicting future pipe breaks (Snider and McBean 2018; Winkler et al. 2018).

### 3.1. Age and Previous-Break ranking models

Many utilities in North America do not employ advanced statistical models, arguing that they lack the necessary data required to calibrate the models (Rao and Francis 2014). Instead, these utilities often rely on simplistic models that base their prioritization for pipe replacement by rank-ordering the pipes from most likely to break to least likely to break, without requiring model calibration. The two most common approaches for ordering these pipes are based on either the age of the pipe or the number of previous breaks (Kirmeyer, Richards, and Smith 1994). This paper compares both of these simple pipe break ranking models.

The Age-model simply ranks all pipes in the test dataset based on their age. The model assumes 'the older the pipe, the shorter time-to-next-break'. To compare the accuracy of the Age-model, to other models, the Age-model ranks the pipes based on their age at last-recorded break within the test dataset. If no break is recorded in the test dataset, the pipe's age is calculated as the age of the pipe at the start of the break records. The accuracy of the model is then calculated by comparing the age ranking to the rank of the earliest time-to-next break observed in the test dataset.

The Previous-Break model counts the number of recorded breaks a pipe has already experienced within the training dataset timeframe. The pipes are then ranked from the largest number of previous breaks to the least number of previous breaks. The pipe with the largest number of previous breaks is considered to have the shortest time-to-next-break within the test dataset.

Since the Previous-Break models and Age-models each rely on a single variable (the number of previous breaks or the age of the pipe), a dataset with limited/missing input variables (i.e. missing pipe diameter, pipe length, soil conditions, etc.) will not impact these models' accuracy. Also, since these simple ranking models do not require calibration, the number of pipes within the training dataset will not impact their accuracy. However, limiting the break history records may affect these models since shorter break records would eliminate some of the observed breaks, impacting both variables of interest (the number of previously recorded breaks or the age of the pipe at last-recorded break).

### 3.2. Extreme gradient boosting decision tree machine learning model

Numerous machine learning models have been developed in order to predict pipe breaks, such as artificial neural network, XGboost, Random Forest, support vector machines and evolutionary polynomial regression. Although the algorithms vary between models, the underlying process is similar, with all machine learning models adopting a data-driven approach to identify complex relationships between input variables and the prediction.

The machine learning model developed for this study relies on the extreme gradient boosting algorithm, or XGBoost (Chen et al. 2018). This algorithm has strong predictive capabilities, is



not susceptible to outliers, and has a relatively short training time. Pipe break prediction models developed using the extreme boosting decision tree algorithm have been shown to outperform other machine learning models such as ANN and Random Forest (Snider and McBean 2018; Winkler et al. 2018). Various utilities have adopted similar decision tree ensembles to predict pipe failure (Fracta 2019). For these reasons, the XGBoost model was chosen to be investigated in this research.

XGBoost is an ensemble machine learning algorithm, built by combining numerous decision trees. A decision tree is a simple, but effective, supervised machine learning model, which creates a prediction based on a chain of binary splits. However, decision trees are often associated with high levels of variance (Hastie et al. 2005). 'Boosting' is a common method that combines numerous decision trees in order to lower the variance within the model and improve overall accuracy.

Using pipe break as an example, the XGBoost algorithm begins with an initial prediction of time to next break by calculating the average time to next break for all pipes within the training datasets. Next, the residuals are calculated via the derivative of the loss function and a tree is built to predict these residuals. New residuals are then calculated, and another decision tree is built. This process is repeated adding all additional trees together until residuals no longer contribute a significant improvement in the overall prediction.

This paper develops the XGBoost pipe break prediction model using the xgboost library developed by (Chen et al. 2018) employed via computer programming language R (R Core Development Team 2017).

### 3.3. Weibull proportional hazard survival analysis model

The Weibull proportional hazard (Weibull PH) model is a fully parametric model that has been shown to outperform other probabilistic pipe break prediction models, such as a Cox-proportional hazard model and Poisson regression (Kimutai et al. 2015). The time to break for the Weibull PH model is represented by the following equations:

$$\log(T) = B'x + \sigma\epsilon \quad (1)$$

$T$  represents the time-to-break (in years),  $x$  represents a vector of input variables (such as pipe length and pipe diameter),  $B$  represents the weights associated with each input variable,  $\epsilon$  follows extreme minimum value distribution  $G(0, \sigma)$  and  $\sigma$  is a scale parameter for the Weibull distribution (Hosmer, Lemeshow, and Sturdivant 2013).

Frequently, survival analysis models are represented in terms of the probability of surviving to time  $t$ , or survival curve. The survival curve for the Weibull PH model is demonstrated in Equation (2), where  $S(t)$  represents the probability of a pipe not experiencing the break before time  $t$ .

$$S(t) = \exp \left[ - \left( \frac{t}{\exp(B'x)} \right)^{\frac{1}{\sigma}} \right] \quad (2)$$

The weights associated with each input variable ( $B$ ) and the Weibull scale parameter  $\sigma$  can be estimated via the maximum likelihood estimation method. The likelihood function for a general survival model is:

$$L = \prod_n^{i=1} L_i = \prod_i h(t_i)^{d_i} S(t_i) \quad (3)$$

Here,  $n$  represents the number of units (or pipes in this case),  $h(t)$  represents the hazard curve (or instantaneous break rate),  $d_i$  is an indicator variable highlighting whether or not the pipe is right-censored and  $S(t)$  represents the survival curve. For pipes with censored events ( $d_i = 0$ ), Equation (3) becomes only a function of the survival curve. This is because, for a right-censored event, we are only certain that the subject survived past time  $t$ .

By substituting the specific hazard and survival curve for the Weibull PH model, the likelihood function becomes:

$$L(B, \sigma | y_i) = \prod_n^{i=1} \left\{ \frac{1}{\sigma} \left( \frac{y_i - B'x}{\sigma} \right) \exp \left[ - \exp \left( \frac{y_i - B'x}{\sigma} \right) \right] \right\} \left\{ \exp \left[ - \exp \left( \frac{y_i - B'x}{\sigma} \right) \right] \right\} \quad (4)$$

In Equation (4),  $y_i$  refers to the observed time-to-break (or end of observation) for pipe  $i$ .

Both the maximum likelihood estimation of the weights ( $B$ ) and the Weibull scale parameter ( $\sigma$ ) are calculated using the survival function in R from the survival library (Therneau 2015).

The Weibull PH model predicts the probability of survival for each pipe, for various times, ranging from 100% survival to 0% survival. In order to provide a similar comparison with other models, a single time to next break must be selected. The obvious choice is the median survival time ( $t_{i(50)}$ ), or the time at which the probability of survival is 50%. An estimate of the median survival time for each individual pipe is:

$$t_{i(50)} = \log(2)^{\sigma} \exp(B'x_i) \quad (5)$$

The median survival time is used throughout the rest of this paper to estimate the time-to-next break for the Weibull PH model.

## 4. Model simulations

All pipe break models investigated in this paper rely, to various degrees, on historical databases to inform their predictions. The size of the historical database influences the accuracy of the models. To quantify this impact, various sizes of historical data are used to develop the four pipe break models. This paper investigates three variables that are believed to be crucial in developing an accurate pipe break model: (1) the number of pipes included in the training dataset; (2) years of historical break records within the training dataset; and (3) the various input variables included in each model. The subsequent sections describe how these three variables were adjusted and how the impacts of these variations are quantified.

### 4.1. Number of pipes within the training data

The data used in this case study are from a large utility in North America, with various levels of cast iron, ductile iron, asbestos cement and PVC pipes. The models developed as reported herein focus only on cast iron pipes since this pipe material contains the largest number of pipes (20,799) with the longest break records (1960–2016).

Eight different sample sizes of pipes were randomly selected from the complete cast iron dataset (20,000, 15,000, 10,000, 5,000, 2,000, 1,000, 500, and 250 pipes), to ensure unbiased representation of the data. This random selection process was repeated 10 times and the average results are reported throughout this paper.

#### 4.2. Limiting the years of break records in training data

The historical break data used herein contains break records from 1960 to 2016, over 55 years of pipe break records. Six sample sizes were created using different numbers of break record years: 1960–2006; 1970–2006; 1980–2006; 1990–2006; 1996–2006; 2001–2006. Break records for the period 2006–2016 were removed from all training datasets and utilized as an independent test dataset to assess the accuracy of the model.

Forty-eight different training sets were created by combining the six sample sizes of break record years (1960–2006; 1970–2006; 1980–2006; 1990–2006; 1996–2006; 2001–2006) with eight sample sizes of pipes (20,000, 15,000, 10,000, 5,000, 2,000, 1,000, 500, and 250 individual pipes). Developing models with these different training datasets allows investigation of the role which historical data plays in building an accurate pipe break prediction model.

#### 4.3. Modelling the impacts of limited input variables

Relevant input variables, or covariates, help build a more accurate statistical model by providing additional information pertaining to pipe breakage. These input variables can be numerous and may include information such as installation year, average pressure, pipe length, and diameter. Input variables are often stored in numerous databases and collecting and organizing this information into a single dataset that can be used to calibrate the pipe break prediction model can be difficult. Furthermore, some utilities are missing entire databases. Therefore, the purpose of this analysis is to determine how the presence/absence of certain input variables may affect the accuracy of each pipe break prediction model.

To model the impact of missing variables, the input variables are grouped into similar categories that often represent a distinct database. Therefore, a utility would be able to use this simulation to identify the impact of missing that database.

The categories of data and their associated input variables are listed in Table 1.

#### 4.4. Assessing the accuracy of each model

All models were assessed on their ability to predict the time-to-next break. A single test dataset was used which contained break records for all 20,799 cast iron pipes that occurred

**Table 1.** Grouped variables for limited attributes simulation.

Category	Variables
Pipe Attributes	Diameter, Length
Location Attributes	PSI, Soil, District, Latitude/Longitude
Rehabilitation Attributes	Cathode, Cement Mortar Lining

between the years 2006–2016. The time-to-next break was calculated as the difference, in years, between the break that occurred during the test dataset, and the previous break date for each individual pipe. If no break occurred within the test set for a certain pipe, a censored event was considered at the end of the test dataset (12/30/2016). For pipes with no previous breaks, the time-to-next break in the test dataset was calculated as break date in the test dataset, minus the start of the break records, or the install year, which ever occurred most recently. If multiple breaks occurred for a single pipe within the test set, each additional break was treated as a separate event/record. For these records, the time-to-next break was still calculated as the years from last-recorded break, and appropriate attributes were updated (age of last break, number of previously recorded breaks, etc.). Table 2 provides several examples of these calculations.

A single test set was used for all models; however, the test dataset did change depending on the years of break record year used in the training dataset. Specifically, as years of break records were reduced, the numbers of pipes with previously recorded breaks were also reduced.

#### 5. Concordance index

The Concordance index (or C-index) performance indicator was selected to compare the accuracies of all models developed. The C-index is a ranking metric that incorporates right-censored data and represents the level of concordance, or agreement in order, between the rank provided by the model and the order of observed time-to-next break within the test set.

The algorithm used to calculate the C-index is as follows:

- (1) Create all possible pairwise comparisons for the observed time-to-next break. Omit pairs where the first event is censored, or the first event (time-to-next break) is larger than the second event.
- (2) For all valid observed time-to-next break pairs (i.e. pairs where  $y_1 > y_2$  and  $y_1$  is not censored), test whether the corresponding predictions are concordant, i.e.  $\hat{y}_1 > \hat{y}_2$ , if so, count as 1. If  $\hat{y}_1 = \hat{y}_2$  count as 0.5. If the pairs are not in the same order – count as 0.

**Table 2.** Calculating time-to-next-break.

Pipe ID	Install Year	Start of Break Records	Previously Recorded Breaks	Previously Recorded Break	Next Recorded Break in Test Dataset (2006–2016)	time-to-next-break (yrs)
A <sub>1</sub>	1 September 1963	1 January 1980	5	8 September 1992	31 December 2009	17.32
A <sub>2</sub>	1 September 1963	1 January 1980	6	31 December 2009	31 December 2012	3
B	1 September 1963	1 January 1980	0	NA	31 December 2009	30
C	1 September 1991	1 January 1980	0	NA	31 December 2009	18.34
D	1 September 1963	1 January 1980	1	8 September 1992	NA	24.32+*

\*+ indicates the time-to-next-break is censored



- (3) Calculate the total C-index by summing all values from step (2) and dividing by the total number of valid pairwise pairs.

$$\hat{C} = \frac{1}{num} \sum_{i: d_i=1} \sum_{j: y_i < y_j} I[\hat{y}_i > \hat{y}_j] \quad (7)$$

In Equation (7),  $\hat{C}$  represents the C-index,  $num$  denotes the number of all comparable pairs (i.e. where pipe  $i$  is uncensored, and pipe  $j$  has observed event greater than  $i$ ),  $y$  is the observed time-to-next-break,  $\hat{y}$  represent the models prediction or rank of next break, and  $I[*]$  is an indicator function. Therefore, this metric includes censored data, creating ranked pairs where the observed, uncensored events occur before the observed censored event.

Since the C-index is a ranking metric it enables comparisons between the machine learning and probabilistic models with the two simplistic ranking models. The maximum value for the C-index is 1, which suggests the rank of all observed events is the same as the rank predicted by the model, and a C-index value of 0.5 indicates the prediction model is no better than chance. Example calculations of the C-index for each model are located in the [Appendix A](#) and an in-depth description of the C-index can be found in Harrell, Lee, and Mark (1996).

## 6. Case study

All models in this paper were trained and tested using pipe break records from a single, large, Canadian water distribution system. This distribution system provided break records from 1960–2016 for over 30,000 pipes.

To keep this paper concise, all of the models were employed using a single pipe material – cast iron. Cast iron is the obvious pipe material choice for this investigation since it has the largest number of pipes (20,799) and the largest number of breaks (54,789), allowing for a greater investigation into how the size of historical data affects pipe break prediction's accuracy. Furthermore, cast iron pipes are often considered the primary material of concern for most utilities in North America (AWWA 2012; Folkman 2018).

The data distribution of the important cast iron variables is visualized in [Figure 3](#) and the complete list of input variables are outlined in [Table 3](#). It is important to note that the dataset used in this case study is from a real utility, and therefore is subjected to errors that are present in all pipe break datasets. Data cleaning was used to remove some erroneous data (i.e. removing pipe breaks that occurred before pipe installation, removing pipes missing input variables, and/or removing pipes with chosen pipe material corresponding to incorrect construction year). However, some errors are likely to remain within the dataset and while there may have been some impacts on the prediction accuracy of each model, the magnitudes are considered as minimal for the purposes of this research.

The attributes outlined in [Table 3](#) are the complete list of input variables used to train the statistical models. The latitude and longitude of a pipe's centroid were found to be a valuable input variable for the XGBoost machine learning model. Since the XGBoost model is a black box, it is difficult to determine

how the model includes these variables to improve pipe break predictions. However, researchers have identified similar location attributes as useful pipe break predictors since they can identify areas with increased break rates, i.e. break clusters (Oliveira et al. 2011; Kleiner and Rajani 2012).

The latitude and longitude of pipe centroid were not included in the Weibull PH model since it was found to decrease the model's pipe break prediction accuracy. This result was expected since the Weibull PH model explicitly defines a log-linear relationship between the input parameter and the time-to-break prediction. The latitude and longitude of the pipe centroid are not expected to have a log-linear relationship with pipe breaks and therefore, including these variables does not increase the prediction accuracy for the Weibull PH model.

## 7. Results

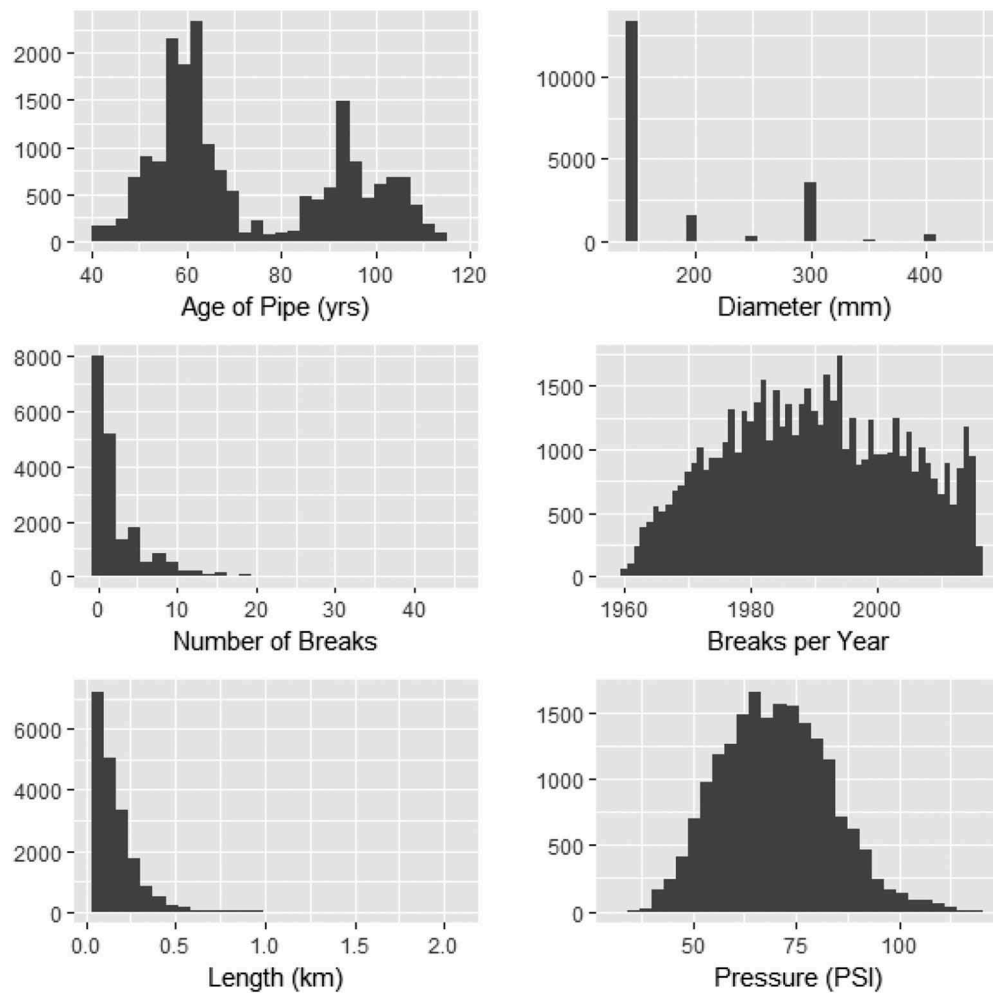
After training the pipe break models using various quantities of historical data, the models' predictions were tested on a 10-year independent test dataset. The accuracy of these predictions was assessed using the C-index performance indicator. The following sections highlight how the C-index, which is a proxy for 'accuracy', is affected by limited pipes within the training dataset, limited break years and limited input variables.

### 7.1. Impact of number of pipes and break years on pipe break prediction

The role which 'number of pipes' and 'break years' play in building an accurate prediction model is investigated congruently, since both these variables impact the number of instances the models are able to 'learn' from. Defining this relationship will assist utilities in selecting the most appropriate model based on their unique data availability. The result of this investigation is outlined in [Figure 4](#).

[Figure 4](#) demonstrates how the 'number of pipes' and 'years of break records' impacts the C-index for various pipe break models. The figure highlights several interesting relationships between data availability and model accuracy:

- **All models are negatively impacted by limiting the break history records.** The accuracy of the statistical models were expected to decrease as the break history was reduced, since the models have less pipe break data to 'learn from', and calibrate the model. The simple Age and Previous-Break ranking models do not rely on a training dataset, but since years of break records impact the testing dataset (by reducing the number of previous breaks or effecting the pipes previous break age), the simplistic models' accuracy is also impacted.
- **Age is a poor indicator of time-to-next break prediction.** The Age-based ranking model performed the worst of the four pipe break prediction models, regardless of break years or numbers of pipes within the training dataset. These findings demonstrate that utilities should avoid using age to predict which pipe will break next.
- **With larger training datasets, XGBoost machine learning pipe break prediction model outperformed all other models.** The XGBoost machine learning model



**Figure 3.** Histograms detailing in-service cast iron pipes for utility described in the case study.

**Table 3.** Case study: complete list of cast iron pipe attributes.

Variable	Range of Values	Description
District	1,2,3,4	district the pipe segment is located in
Diameter (mm)	150–450	diameter of pipe segment
Length (m)	50–2034	length of pipe segment
Cement Mortar Lining	0,1	indicating whether pipe has received mortar lining at time of prediction
Cathode Protection	0,1	indicating whether pipe has received cathodic protection at time of prediction
Latitude	Removed for Anonymity	decimal degrees for centroid of pipe segment, not included in Survival Analysis Model
Longitude	Removed for Anonymity	decimal degrees for centroid of pipe segment, not included in Survival Analysis Model
Pressure (psi)	34–120	Static pressure from nearest hydrant measurement
Soil	clay, sand, bedrock, glacial till	Soil classification
Previous Recorded Breaks	0–45	number of previous breaks recorded on pipe segment
Age of pipe at last break (yrs)	(0–56)	number of years between last break and start of break records (or install year, whichever is more recent)
Age of pipe at second last break (yrs)	(0–56)	number of years between second last break and start of break records (or install year, whichever is more recent)
Age of pipe at third last break (yrs)	(0–56)	number of years between third last break and start of break records (or install year, whichever is more recent)
Years without Break Records	(0–101)	number of years between start of break records and install year

outperforms Age and Previous-Break ranking models and the Weibull PH probabilistic model when 2,000 or more pipes and more than five years of break records are available. The improvement in accuracy becomes more noticeable as break years and number of pipes increase. This suggests these larger datasets provide sufficient training

opportunities for the machine learning model to identify important relationships between the input variables and the time-to-next-break.

- **Previous-Break ranking model performed consistently well for all datasets.** The Previous-Break ranking model achieved satisfactory results for all datasets and

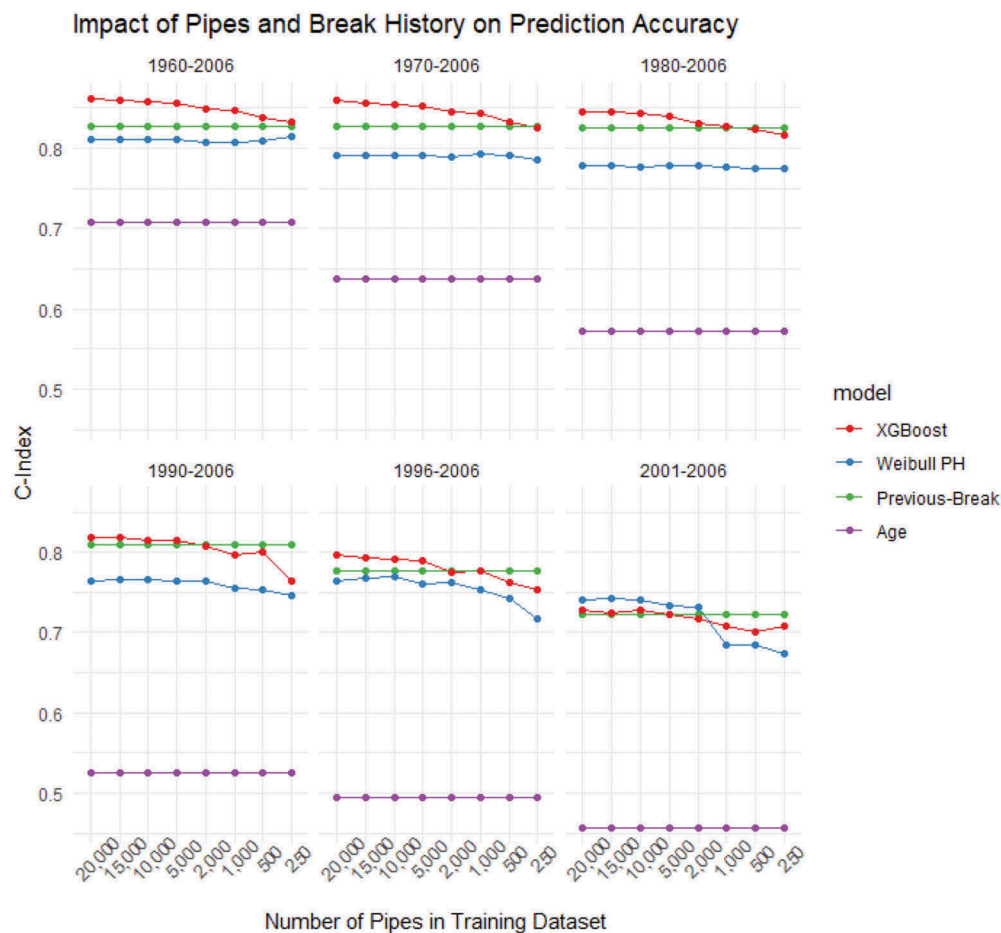


Figure 4. The effects of limited break records and number of pipes on pipe break prediction accuracy.

outperformed all other models when pipe break data was limited (less than ten years of break records and less than 2,000 pipes).

Overall, Figure 4 demonstrates the important role break records and numbers of pipes play in developing accurate pipe break prediction models. These findings indicate the XGBoost machine learning model has real utility when the training dataset contains 2,000 or more pipes and more than 5 years of break records. Also worth noting is that the Previous-Break ranking model received a consistently high C-index even outperforming all other models when the training dataset is severely limited in size. However, it is important to be aware of some of the drawbacks associated with the Previous-Break ranking model, when compared to statistical models such as the Weibull PH and the XGBoost model that are not evident in Figure 4. These drawbacks include the large number of 'ties' (pipes with the same number of previous breaks and thus equal likelihood of failure) associated with the Previous-Break prediction model (especially with small datasets) and only ranking the likelihood of failure, instead of predicting the time of failure, or break.

Whereas the statistical models predict a continuous variable (time-to-next-break) the Previous-Break ranking model relies on

the discrete number of previous breaks recorded for each pipe segment. This results in many pipes having the same rank, since the number of previous breaks are equal. These 'ties' may make it difficult for the utility to prioritize pipe replacement, since the previous-break model is unable to distinguish which of these pipes will break first. For example, if the utility in this case study has five years of break records and decides to replace 20 of the worst pipes then the first 12 pipes are easily selected, since they all have greater than four previously recorded breaks. However, the remaining eight pipes will need to be selected from 36 pipes which all have four previously recorded breaks. Making this selection may be difficult. If instead, the utility adopts a statistical model, they would be able to directly select the top 20 pipes with the shortest predicted time to next break, since the prediction is a continuous variable with essentially zero ties.

The second disadvantage associated with the Previous-Break model is that it simply ranks the pipes and provides no indication as to when the pipe is most likely to fail. If utilities require an estimated time-to-next break for their pipe replacement plan, they should not consider either the Previous-Break or Age simplistic ranking models, instead of relying on statistical models that explicitly predict time to next break for each pipe.

## 7.2. Impact of limited array of input variables on models' accuracy

Utilities not only experience limitations of recorded years of pipe breaks, and a limited number of pipes to calibrate their pipe break prediction model, they also experience incomplete data with regards to input variables, such as missing pipe diameter, pipe length, average pressure, and so on. Therefore, it is important to investigate how various pipe break prediction models are impacted by incomplete input variables.

To investigate the impact of missing input variables, each model was built with various input variables removed and their overall accuracy assessed with the C-index. The input variables removed were grouped into specific categories representing attributes that are often found in separate databases and therefore, sometimes difficult to assemble. For example, records of cement mortar relining and cathode protection are typically located in rehabilitation databases. The grouped variables as assumed for purposes of this paper, are outlined in Table 1 with the complete list of input variables outlined in Table 3. Figure 5 depicts the accuracy of each model with certain groups of variables removed.

Figure 5 demonstrates several key findings:

- **The XGBoost machine learning model is robust against missing input variables.** The maximum reduction in the machine learning model's C-index (reduced by 0.01) is

noticed when location attributes (latitude and longitude of pipe centroid) are removed. This is a relatively minor reduction in overall accuracy and may be due to the machine learning model's ability to identify important non-linear relationships between covariates and the output. If one variable is removed, the machine learning model may be able to identify complex relationships between the remaining variables and represent a portion of the missing signal associated with the removed variable.

- **Pipe attributes (length and diameter) are the most important input variables when building the probabilistic pipe break prediction model.** The Weibull PH survival analysis model experienced a substantial decrease in C-index (overall C-index reduced by 0.15) when pipe attribute input variables were removed. This suggests that a strong log-linear relationship exists between the pipe attributes and future pipe breaks. The importance of the pipe attribute variables have been suggested by other various researchers (Hu and Hubble 2007; Kimutai et al. 2015).
- **Simple Ranking Models (Age and Previous-Break) are not impacted by limited input variables.** The Age and Previous-Break models are not impacted when input variables are reduced since these models only rely on a single input variable to rank the likelihood of future pipe break.

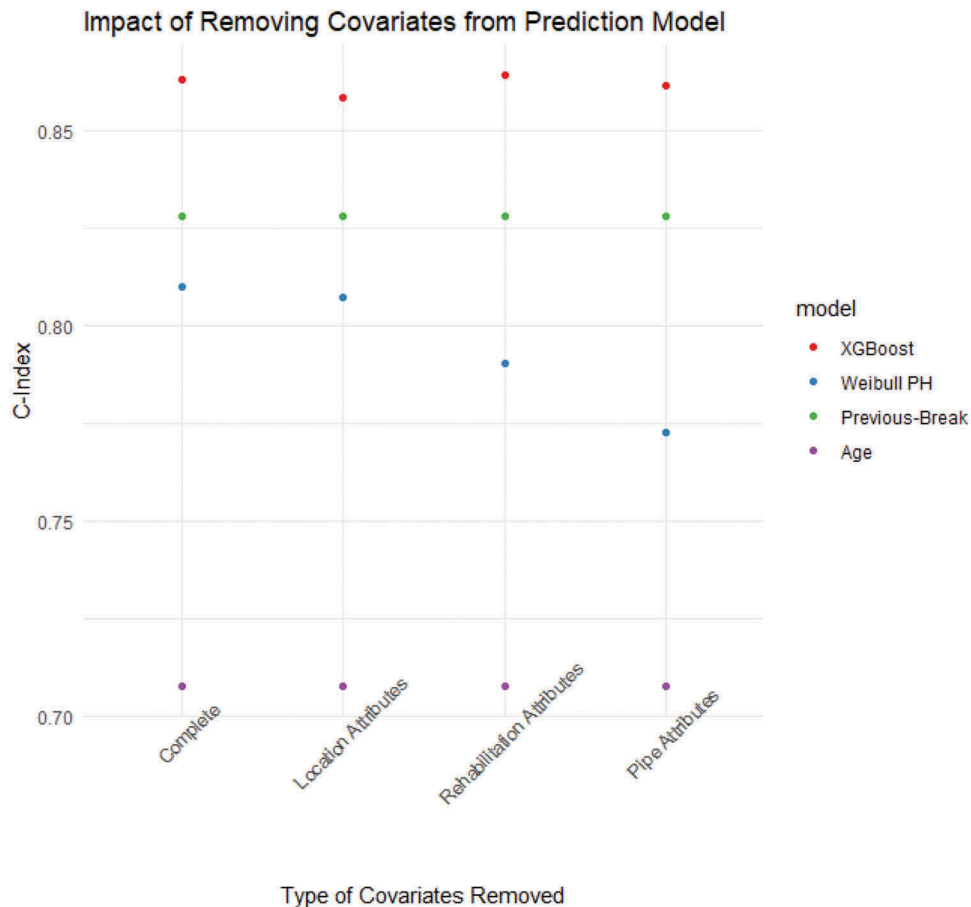


Figure 5. Impact of limited input variables on models' accuracy.

Overall, **Figure 5** demonstrates that even with key input variables missing, both the probabilistic and the machine learning pipe break models outperform the Age pipe break ranking model. Furthermore, the XGBoost machine learning model is not substantially impacted by missing input variables. Therefore, utilities concerned with missing key input variables may consider adopting the XGBoost machine learning algorithm to develop accurate pipe break predictions.

## 8. Conclusion

Many of the water distribution pipes installed in North America are approaching their expected end-of-life, and pipe breaks are increasing. To address this, many pipe break prediction models have recently been developed to help utilities prioritize pipe replacement and reduce future pipe breaks. However, many utilities continue to rely on simple age-based or numbers of previous break ranking models, with many citing that they lack sufficient historical data to build and calibrate more complex statistical models (Kirmeyer, Richards, and Smith 1994). The specific role which historical databases play in building various types of pipe break prediction models remains unclear. Therefore, the purpose of this paper was to examine how various pipe break prediction models are impacted by varying magnitudes of historical data.

This paper compared the accuracy of four models, Weibull PH probabilistic model, XGBoost machine learning model, simplistic Age ranking model and simplistic Previous-Break ranking model, with varying degrees of break years, pipes, and input variables. The results of this comparison lead to four main conclusions:

- (1) Simplistic Age-based pipe break model is a poor indicator of a next-pipe break. Instead, utilities should rely on statistical models, or simple Previous-Break ranking model, even when datasets are limited.
- (2) Simplistic Previous-Break model performs well in ranking the likelihood of next break under all data limitations and outperforms all other models when number of break years and number of pipes are limited. However, it may be difficult for a utility to implement the previous-break model due to the model not distinguishing between pipes with equal numbers of previous breaks, and the model not providing an estimated time to next break.
- (3) The XGBoost machine learning algorithm outperforms all other models when data records containing more than 5 years of break records and 2,000 or more pipes are used to calibrate the model.
- (4) The accuracy of the XGBoost machine learning model is not heavily impacted by limited input variables.

Overall, the four models selected for this paper represent three unique types of pipe break prediction models; simplistic, probabilistic and machine learning. The models were selected based on their perceived accuracy as well as the popularity of use. Although the accuracy of the results will depend on the specific model selected, and the dataset used, the conclusion drawn from this research may provide insight into how different simplistic, probabilistic and machine learning models are impacted

by limited break data, and help further guide utilities in selecting the appropriate model based on their data availability.

This paper was limited to a single case study, focusing only on cast iron pipes. Future research should look to corroborate these results with other utility break records and different pipe materials. This comparison may provide insights into how inaccuracies within the raw data and, in particular, the accuracy of pipe installation records and pipe break dates, may impact the accuracy of each model. Comparing model accuracy across a number of utilities with various pipe materials, number of pipes and range of pipe break records, would provide more knowledge on the robustness of each pipe break prediction model.

## Acknowledgements

This research was funded by the Natural Sciences and Engineering Research Council (NSERC). The authors are grateful for the help and data provided by the utility described in the case study of this report. In this research, data were processed with R-studio using R-language.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

## ORCID

Brett Snider  <http://orcid.org/0000-0003-4883-6045>

## Data availability statement

All data used during the study are confidential in nature and cannot be provided by agreement with the municipalities due to their concern with the security of their distribution system.

## References

- Achim, D., F. Ghotb, and K. J. McManus. 2007. "Prediction of Water Pipe Asset Life Using Neural Networks." *Journal of Infrastructure Systems* 13 (1): 26–30. doi:[10.1017/CBO9781107415324.004](https://doi.org/10.1017/CBO9781107415324.004).
- Ahn, J., S. Lee, G. Lee, and J. Koo. 2005. "Predicting Water Pipe Breaks Using Neural Network." *Water Science & Technology: Water Supply* 5 (3–4): 159–172.
- Alvisi, S., and M. Franchini. 2010. "Comparative Analysis of Two Probabilistic Pipe Breakage Models Applied to a Real Water Distribution System." *Civil Engineering and Environmental Systems* 27 (1): 1–22. doi:[10.1080/10286600802224064](https://doi.org/10.1080/10286600802224064).
- Asnaashari, A., E. A. McBean, B. Gharabaghi, and D. Tutt. 2013. "Forecasting Watermain Failure Using Artificial Neural Network Modelling." *Canadian Water Resources Journal* 38 (1): 24–33. doi:[10.1080/07011784.2013.774153](https://doi.org/10.1080/07011784.2013.774153).
- Asnaashari, A., E. A. McBean, I. Shahrou, and B. Gharabaghi. 2009. "Prediction of Water Main Failure Frequencies Using Multiple and Poisson Regression." *Water Science and Technology Water Supply* 9 (1): 9–19. doi:[10.2166/ws.2009.020](https://doi.org/10.2166/ws.2009.020).
- AWWA. 2012. "Buried No Longer: Confronting America's Water Infrastructure Challenge." 37.
- Berardi, L., O. Kapelan, O. Giustolisi, and D. Savic. 2008. "Development of Pipe Deterioration Models for Water Distribution Systems Using EPR." *Journal of Hydroinformatics* 10 (2): 113–126. doi:[10.2166/hydro.2008.012](https://doi.org/10.2166/hydro.2008.012).



- Boxall, J. B., A. O'Hagan, S. Pooladsaz, A. J. Saul, and D. M. Unwin. 2007. "Estimation of Burst Rates in Water Distribution Mains." *Institution of Civil Engineers Water Management* 160 (2): 73–82. doi:10.1680/wama.2007.160.2.73.
- Chen, T., T. He, M. Benesty, V. Khotilovich, and Y. Tang. 2018. "Xgboost: Extreme Gradient Boosting."
- Clark, R., J. Carson, R. Thurnau, R. Krishnan, and S. Panguluri. 2010. "Condition Assessment Modeling for Distribution Systems Using Shared Frailty Analysis." *Journal-American Water Works Association* 102 (7): 81. doi:10.1002/j.1551-8833.2010.tb10151.x.
- Dawood, T., E. Elwakil, H. Novoa, and J. Delgado. 2019. "Water Pipe Failure Prediction and Risk Models: State-of-the-Art Review." *Canadian Journal of Civil Engineering*, no. ja. doi:10.1139/cjce-2019-0481.
- Debón, A., A. Carrión, E. Cabrera, and H. Solano. 2010. "Comparing Risk of Failure Models in Water Supply Networks Using ROC Curves." *Reliability Engineering and System Safety* 95 (1): 43–48. doi:10.1016/j.res.2009.07.004.
- EPA. 2010. "Control and Mitigation of Drinking Water Losses in Distribution Systems."
- Folkman, S. 2018. "Water Main Break Rates in the USA and Canada: A Comprehensive Study." *Mechanical and Aerospace Engineering Faculty Publications*, no. March: 1–49.
- Fracta. 2019. "Machine Learning Comes of Age in the Water Industry - A White Paper with Business Case Studies." Redwood City.
- Fuchs-Hanusch, D., B. Kornberger, F. Friedl, and R. Scheucher. 2012. "Whole of Life Cost Calculations for Water Supply Pipes." *Water Asset Management International* 8 (2): 19–24.
- Harrell, F., K. Lee, and D. Mark. 1996. "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors." *Statistics in Medicine* 15 (4): 361–387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
- Hastie, T., R. Tibshirani, K. Friedman, and J. Franklin. 2005. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd ed. New York, NY: Springer.
- Hatler, D. 2019. "The Accuracy and Business Value of Fracta Machine Learning Drinking Water Main Condition Assessment." <https://blog.fracta.ai/condition-assessment-accuracy-value>
- Hosmer, D., S. Lemeshow, and R. Sturdivant. 2013. *Applied Logistic Regression*. Hoboken, New Jersey: John Wiley & Sons.
- Hu, Y., and D. Hubble. 2007. "Factors Contributing to the Failure of Asbestos Cement Water Mains." *Canadian Journal of Civil Engineering* 34 (5): 608–621. doi:10.1139/j06-162.
- Jafar, R., I. Shahrour, and I. Juran. 2010. "Application of Artificial Neural Networks (ANN) to Model the Failure of Urban Water Mains." *Mathematical and Computer Modelling* 51 (9–10): 1170–1180. doi:10.1016/j.mcm.2009.12.033.
- Kimutai, E., G. Betrie, R. Brander, R. Sadiq, and S. Tesfamariam. 2015. "Comparison of Statistical Models for Predicting Pipe Failures: Illustrative Example with the City of Calgary Water Main Failure." *Journal of Pipeline Systems Engineering and Practice* 6: 4. doi:10.1061/(ASCE)PS.1949-1204.0000196.
- Kirmeyer, G. J., W. Richards, and C. D. Smith. 1994. *Assessment of Water Distribution Systems and Associated Research Needs*. Denver, Colorado: AWWA.
- Kleiner, Y., and B. Rajani. 2000. "Considering Time-Dependent Factors in the Statistical Prediction of Water Main Breaks." *American Water Works Association Infrastructure Conference Proceedings*, Baltimore, Maryland, 1–12.
- Kleiner, Y., and B. Rajani. 2001. "Comprehensive Review of Structural Deterioration of Water Mains: Statistical Models." *Urban Water* 3 (3): 131–150. doi:10.1016/S1462-0758(01)00033-4.
- Kleiner, Y., and B. Rajani. 2002. "Forecasting Variations and Trends in Water-Main Breaks." *Journal of Infrastructure Systems* 8 (4): 122–131. doi:10.1061/(ASCE)1076-0342(2002)8:4(122).
- Kleiner, Y., and B. Rajani. 2012. "Comparison of Four Models to Rank Failure Likelihood of Individual Pipes." *Journal of Hydroinformatics* 14 (3): 659–681. doi:10.2166/hydro.2011.029.
- Kumar, A., S. Rizvi, B. Brooks, R. Vanderveld, K. Wilson, C. Kenney, S. Edelstein, et al. 2018. "Using Machine Learning to Assess the Risk of and Prevent Water Main Breaks." In *SIGKDD'18*, 1–9. London, UK. doi:10.1145/nmnnnnnn.nnnnnnnn.
- Laucelli, D., B. Rajani, Y. Kleiner, and O. Giustolisi. 2014. "Study on Relationships between Climate-Related Covariates and Pipe Bursts Using Evolutionary-Based Modelling." *Journal of Hydroinformatics* 16 (4): 743. doi:10.2166/hydro.2013.082.
- Le Gat, Y., and P. Eisenbeis. 2000. "Using Maintenance Records to Forecast Failures in Water Networks." *Urban Water* 2 (3): 173–181. doi:10.1016/S1462-0758(00)00057-1.
- Nishiyama, M., and Y. Filion. 2014. "Forecasting Breaks in Cast Iron Water Mains in the City of Kingston with an Artificial Neural Network Model." *Canadian Journal of Civil Engineering* 41 (10): 918–923. doi:10.1139/cjce-2014-0114.
- Oliveira, D. P., D. Neill, J. Garrett, and L. Soibelman. 2011. "Detection of Patterns in Water Distribution Pipe Breakage Using Spatial Scan Statistics for Point Events in a Physical Network." *Journal of Computing in Civil Engineering* 25 (1): 21–30. doi:10.1061/(ASCE)CP.1943-5487.0000079.
- Park, S., H. Jun, N. Agbenowosi, B. J. Kim, and K. Lim. 2011. "The Proportional Hazards Modeling of Water Main Failure Data Incorporating the Time-Dependent Effects of Covariates." *Water Resources Management* 25 (1): 1–19. doi:10.1007/s11269-010-9684-y.
- Park, S., J. Kim, A. Newland, B. J. Kim, and H. Jun. 2008. "Survival Analysis of Water Distribution Pipe Failure Data Using the Proportional Hazards Model." *World Environmental and Water Resources Congress*, Ahupua'a.
- Park, S., J. Kim, A. Newland, and H. Jun. 2007. "A Methodology to Estimate Economically Optimal Replacement Time Interval of Water Distribution Pipes." *Water Science & Technology: Water Supply* 7 (5–6): 149–155.
- R Core Development Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria.
- Rajani, Y., and B. Kleiner. 2001. "Comprehensive Review of Structural Deterioration of Water Mains: Statistical Models." *Urban Water Journal* 3 (3): 131–150. doi:10.1016/S1462-0758(01)00033-4.
- Rao, V. M., and R. A. Francis. 2014. "Investigating the Role of Statistical Models in Water Distribution Asset Management: A Semi-structured Interview Approach." *Probabilistic Safety Assessment and Management PSAM 12*, Honolulu, Hawaii, June 2014.
- Renzetti, S., and D. Dupont. 2013. *Buried Treasure: The Economics of Leak Detection and Water Loss Prevention in Ontario*. Rep. No. E.
- Savic, D., O. Giustolisi, and D. Laucelli. 2009. "Asset Deterioration Analysis Using Multi-Utility Data and Multi-Objective Data Mining." *Journal of Hydroinformatics* 11 (3–4): 211–224. doi:10.2166/hydro.2009.019.
- Scheidegger, A., J. P. Leitão, and L. Scholten. 2015. "Statistical Failure Models for Water Distribution Pipes - A Review from A Unified Perspective." *Water Research* 83: 237–247. doi:10.1016/j.watres.2015.06.027.
- Snider, B., and E. A. McBean. 2018. "Improving Time to Failure Predictions for Water Distribution Systems Using Gradient Boosting Algorithm." *WDSA/CCWI Joint Conference Proceedings*, Kingston, Canada.
- St. Clair, A. M., and S. Sinha. 2012. "State-of-the-Technology Review on Water Pipe Condition, Deterioration and Failure Rate Prediction Models!" *Urban Water Journal* 9 (2): 85–112. doi:10.1080/1573062X.2011.644566.
- Tabesh, M., J. Soltani, R. Farmani, and D. Savic. 2009. "Assessing Pipe Failure Rate and Mechanical Reliability of Water Distribution Networks Using Data-Driven Modeling." *Journal of Hydroinformatics* 11 (1): 1–17. doi:10.2166/hydro.2009.008.
- Therneau, T. 2015. "A Package for Survival Analysis in S. Version 2.38."
- Underhill, M. D., and I. Elliott. 2012. "Water and Wastewater Infrastructure." *The Handbook of Infrastructure Investing*, no. March: 63–79. doi:10.1002/9781118268117.ch5.
- Van Buuren, S., H. C. Boshuizen, and D. L. Knook. 1999. "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis." *Statistics in Medicine* 18 (6): 681–694. doi:10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R.
- Vanrenterghem-Raven, A., P. Eisenbeis, I. Juran, and S. Christodoulou. 2004. "Statistical Modeling of the Structural Degradation of an Urban Water Distribution System: Case Study of New York City." *World Water & Environmental Resources Congress 2003*, Pennsylvania, USA.



- Wang, Y., T. Zayed, and O. Moselhi. 2009. "Prediction Models for Annual Break Rates of Water Mains." *Journal of Performance of Constructed Facilities* 23 (1): 47–54. doi:[10.1061/\(ASCE\)0887-3828-\(2009\)23:1\(47\)](https://doi.org/10.1061/(ASCE)0887-3828(2009)23:1(47)).
- Wiesner, R. H., P. M. Grambsch, E. R. Dickson, J. Ludwig, R. L. Maccarty, E. B. Hunter, T. Fleming, L. Fisher, S. Beaver, and N. Larusso. 1989. "Primary Sclerosing Cholangitis: Natural History, Prognostic Factors and Survival Analysis." *Hepatology* 10 (4): 430–436. doi:[10.1002/hep.1840100406](https://doi.org/10.1002/hep.1840100406).
- Wilson, D., Y. Fillion, and I. Moore. 2015. "State-of-the-Art Review of Water Pipe Failure Prediction Models and Applicability to Large-Diameter Mains." *Urban Water Journal* 14 (2): 173–184. doi:[10.1080/1573062X.2015.1080848](https://doi.org/10.1080/1573062X.2015.1080848).
- Winkler, D., M. Haltmeier, M. Kleidorfer, W. Rauch, and F. Tscheikner-Gratl. 2018. "Pipe Failure Modelling for Water Distribution Networks Using Boosted Decision Trees." *Structure and Infrastructure Engineering* 14 (10): 1402–1411. doi:[10.1080/15732479.2018.1443145](https://doi.org/10.1080/15732479.2018.1443145).
- Xu, Q., Q. Chen, and W. Li. 2011. "Application of Genetic Programming to Modeling Pipe Failures in Water Distribution Systems." *Journal of Hydroinformatics* 13 (3): 419–428. doi:[10.2166/hydro.2010.189](https://doi.org/10.2166/hydro.2010.189).
- Xu, Q., Q. Chen, W. Li, and J. Ma. 2011. "Pipe Break Prediction Based on Evolutionary Data-Driven Methods with Brief Recorded Data." *Reliability Engineering and System Safety* 96 (8): 942–948. doi:[10.1016/j.res.2011.03.010](https://doi.org/10.1016/j.res.2011.03.010).

Yamijala, S., S. D. Guikema, and K. Brumbelow. 2009. "Statistical Models for the Analysis of Water Distribution System Pipe Break Data." *Reliability Engineering and System Safety* 94 (2): 282–293. doi:10.1016/j.res.2008.03.011.

## Appendix A. C-index example calculations

The C-index, as described by Harrell, Lee, and Mark (1996), is used throughout this paper to compare accuracies between simplistic ranking models and

**Table A1.** Example observations for C-index calculation.

Pipe ID	Age	Previous Recorded Breaks	Weibull Prediction (yrs)	XGBoost Prediction (yrs)	Observed time-to-next-break (yrs)
Pipe A <sub>1</sub>	29	5	7	17	17.32
Pipe A <sub>2</sub>	46	6	10	5	3
Pipe B	0	0	20	31	30
Pipe C	30	1	40	50	24.32+*
Pipe D	25	1	5	13	15

\*+ indicates the time-to-next-break is censored (didn't occur during test dataset)

statistical models. The following section provides sample calculations for the C-index for each model, using a small dataset.

**Table A1** below highlights six observed events and their prediction or rank by each model. Note not all pipe attributes or input variables that statistical models rely on are included in the table below. In this example Pipe A has two recorded break records within the test data and therefore occurs twice within this table. Pipe A<sub>1</sub> refers to the first break within the test dataset, and Pipe A<sub>2</sub> refers to the second break.

The first step in calculating the C-index is to create all possible pairwise comparisons for the observed time-to-next break, omitting pairs where the first

**Table A2.** Pairwise Comparison.

Pair of Pipe IDs	Pair of Observed Events	Age Pair	Previous Recorded Break Pair	Weibull Prediction Pair	XGBoost Prediction Pair
(A <sub>2</sub> ,D)	(3,15)	(46,25)	(6,1)	(10,5)	(5,13)
(A <sub>2</sub> , A <sub>1</sub> )	(3,17.32)	(46,29)	(6,5)	(10,7)	(5,17)
(A <sub>2</sub> ,C)	(3,24.32+)	(46,30)	(6,1)	(10,40)	(5,50)
(A <sub>2</sub> ,B)	(3,30)	(46,0)	(6,0)	(10,20)	(5,31)
(D,A <sub>1</sub> )	(15,17.32)	(25,29)	(1,5)	(5,7)	(13,17)
(D,C)	(15,24.32+)	(25,30)	(1,1)*	(5,40)	(13,50)
(D,B)	(15,30)	(25,0)	(1,0)	(5,20)	(13,31)
(A <sub>1</sub> ,C)	(17.32,24.32+)	(29,30)	(5,1)	(7,40)	(17,50)
(A <sub>1</sub> ,B)	(17.32,30)	(29,0)	(5,0)	(7,20)	(17,31)
# of Concordant Predictions		6	7.5	7	9

\*Tied event, counts as 0.5.

event is censored, or the first even (time-to-next break) is larger than the second event. The total number of possible pairs for the five events, ignoring pairs where censored events occur first, is five choose 2 (or 5C2) which equals 10. The total number of pairs where a censored event occurs before an uncensored event is only one (Pipe D = 24.32+ and Pipe B = 30). Therefore, the total number of possible pairs for this dataset is 9. All possible observed pairs are listed in **Table A2**, with corresponding prediction pairs.

The next step to calculating the C-index is to determine the number of concordant pairs for each model. A concordant pair is one where the order of the observed events (the observed time-to-next break) follows the same rank or prediction of the model. For the Age and Previous recorded break models, the greater the age, or the larger number of previous breaks, the assumed shorter time-to-next break.

Therefore, for these models the prediction is considered to be concordant, if the age or previous break of the second event is shorter than the first event. For example, the pair A<sub>2</sub> and D are concordant for Age and Previous Break ranking models. Note if the prediction is tied (as is the case for the previous recorded break pair D and C) then the pair is counted as 0.5.

The Weibull PH and XGBoost models both predict time-to-next break. Therefore, these predictions are considered to be concordant if they follow the same order as observed events. Therefore, in **Table A2**, the second predicted event should be greater than the first predicted event.

The C-index for each model in the above example can now be easily calculated by dividing the total number of concordant predictions by the total number of possible pairwise comparisons. Therefore, the C-index for the example outlined in **Tables A1** and **A2** is as follows:

$$C\text{-index}_{\text{Age}} = 6/9 = 0.67$$

$$C\text{-index}_{\text{PreviousBreak}} = 7.5/9 = 0.83$$

$$C\text{-index}_{\text{WeibullPH}} = 7/9 = 0.78$$

$$C\text{-index}_{\text{XGBoost}} = 9/9 = 1$$