engineering proceedings

MDPI

Proceeding Paper

Developing an Open Repository of Water Main Break Prediction Models in Kitchener †

Fatemeh Boloukasli ahmadgourabi * and Rebecca Dziedzic

Building, Civil and Environmental Engineering Department, Concordia University, Montreal, QC H3G 2W1, Canada; rebecca.dziedzic@concordia.ca

* Correspondence: fatemeh.boloukasliahmadgourabi@mail.concordia.ca

† Presented at the 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024), Ferrara, Italy, 1–4 July 2024.

Abstract: This study presents an open repository of predictive machine learning models for water main breaks as a way to help manage water supply networks proactively. Lack of standardized datasets has been a challenge in previous research, a problem which is addressed in the present study through provision of a benchmark dataset that features pipe dimensions, age, proximity to previous breaks, and climatic variables, among other elements. The repository allows for model testing and comparison with machine learning algorithms such as XGBoost and LightGBM. Implemented in Python and available on GitHub, this project promotes a collaborative approach towards the enhancement of urban infrastructure management through accurate prediction of water main breaks, leading to fewer interruptions in service. Findings show that, while random splits work well in training and testing, their performance is poor when it comes to future prediction. Conversely, time-based splits maintain a good consistency between training and testing phases, but they lack the capacity to predict future periods.

Keywords: water main break; open repository; prediction models

check for updates

Citation: Boloukasli ahmadgourabi, F.; Dziedzic, R. Developing an Open Repository of Water Main Break Prediction Models in Kitchener. *Eng. Proc.* **2024**, *69*, 13. https://doi.org/10.3390/engproc2024069013

Academic Editors: Stefano Alvisi, Marco Franchini, Valentina Marsili and Filippo Mazzoni

Published: 29 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Water supply networks are crucial and costly assets for cities. Interruptions in these systems can not only impact the water supply by causing water loss and contamination, but also damage surrounding infrastructure like sewers, roads, and gas lines, potentially causing significant failures. The United States and Canada have experienced over 2 million water main failures since 2000, resulting in an average of 700 breaks per day, at an annual cost of more than CAD 10 billion [1]. Essential to predicting future water main deterioration is the understanding and modelling of the impacts of various driving factors. These factors are of three types: physical (pipe material, age, length, and diameter), operational (internal water pressure and previous failures), and environmental (dramatic seasonal variation in weather temperature) [2]. Machine learning models have been increasingly employed to predict pipe failure based on these factors, with promising results [3,4].

Previous machine learning studies on the prediction of water main failure rely on distinct case-study systems, and the clean data have not been made available. For example, Fan et al. [3] utilized a multi-source data-aggregation framework incorporating factors of intrinsic, operational, environmental, and societal impacts. Alternatively, Snider and McBean [4] utilized operational and intrinsic data for estimating the time to the next pipe failure, only focusing on ductile iron pipes. On the other hand, Chen [5] et al. aggregate data from multiple sources, utilizing a comprehensive dataset from six U.S. water utilities. This dataset encompassed diverse attributes such as intrinsic factors and historical break records, and was enriched with environmental and demographic data. Omar's [6] study employed three primary datasets from the City of Kitchener, including a water-main assets

inventory, a history of breaks, and road segment data with traffic information. Not all of the code associated with these studies is publicly shared.

These studies generally compare the application of different machine learning models. However, without more insight into the underlying data, the extendibility of the results is unclear. Fan et al. [3] identified LightGBM as the most effective, leveraging its abilities to handle categorical variables and ensure computational efficiency. Conversely, Snider and McBean's investigation highlighted the superiority of gradient boosting algorithms in estimating the time to the next pipe failure [4]. Similarly, Omar et al. [6] pointed to Random Forest as top performer. These variations hint at the complexity of the modeling of water main failures and the influence of data characteristics on model performance.

Furthermore, data availability, as well as the associated cleaning and preparation, have been found to significantly impact results, making the comparison of studies even more complicated. In examining data preparation strategies, there are two major approaches: random splitting and time-based splitting. Random splitting divides the dataset into training and testing sets without considering the temporal order of the data. Time-based splitting organizes data by time periods, typically using the most recent data for testing to simulate the model's performance in predicting future events based on historical records. This approach better matches the real potential applications of these models [7]. Nevertheless, multiple previous studies [4–6] have randomly split their datasets into training and test sets. Although their results appeared promising, these models likely suffer from overfitting, due to the employment of random splitting in a temporal dataset.

To address these challenges, this study aims to develop an open repository that includes predictive machine learning models for water main breaks. By providing a consolidated and consistent dataset, the repository offers a standardized platform for rigorous evaluation of predictive performance. This enhances the reproducibility of research by making data, models, code, and directions available for others to reuse and validate, which is fundamental for the advancement of scientific knowledge and trust in research findings [8].

2. Materials and Methods

An open repository on GitHub houses all project materials, from code to datasets, to ensure transparency and replicability. This repository is structured to ensure ease of access and usability. Data files and scripts are organized with consistent formatting and nomenclature, so researchers interested in replicating or building upon this study can easily utilize them. The entire project is implemented in Python and leverages a benchmark dataset to advance the understanding of the optimal modelling methodologies used for predicting water main breaks. The steps of the project are illustrated in Figure 1.

In this study, three datasets are used: a water main inventory, a history of breaks, and weather data. The first two datasets were collected from the open-data portal of the city of Kitchener [9], and the third from Environment and Climate Change Canada (ECCC) [10]. The original inventory dataset contains fundamental attributes, including pipe length, diameter and material, for pipes installed over the years 1889 to 2023. The break history dataset records instances of pipe failures between 1985 and 2023, and the third dataset contains weather variables such as minimum, maximum, and mean temperatures, along with total rainfall, for the period spanning from 1973 to 2023.

The data cleaning process involved renaming columns and dropping missing values in the break and water main inventory datasets. The weather data were combined to form an integrated file. To enable the training and testing of the models over different time intervals, 5-year intervals were defined for the pipes and the weather data. The datasets were combined by using the intervals and asset ID of the break records and the water main inventory. Then, the resulting dataset was merged with weather data, based on these intervals.

As part of the feature engineering process, new variables were introduced to augment the model's predictive capability, including the cumulative total number of breaks up to the interval just before the current one, as well as up to two intervals before; break rate;

previous break status; the logarithm and square of the pipe's age, the pipe's perimeter and the area of the pipe; diameter to length ratio; break proximity; cumulative cold and hot days; potential evapotranspiration; freezing index; daily mean, minimum and maximum temperature; total rain; and thawing index.

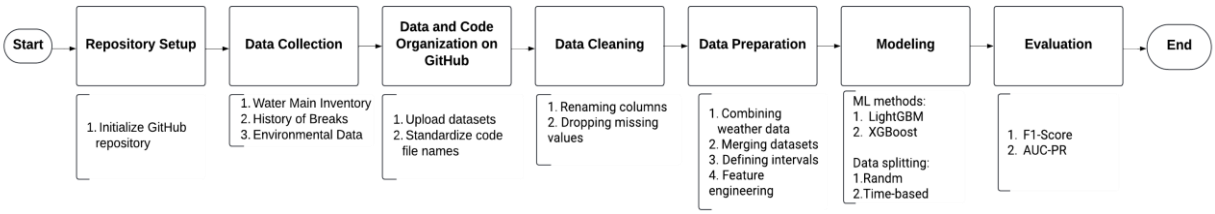

Figure 1. Predictive modeling workflow for water main breaks: from repository setup to evaluation.

Considering the recent research highlighting the superior performance of LightGBM and XGBoost [3,4], these two approaches are adopted herein. Furthermore, because of the potential unreliability of randomly splitting the data [9], random splitting and time-based splitting are compared. In the random split method, first, a subset of data was put aside for the purpose of comparison (2018–2023), and then 20% of the dataset was allocated to the test set, while the remaining 80% formed the training set. Contrastingly, in the time-based split, the data were segmented into a training set spanning from 1984 to 2017 and a testing set covering the period from 2018 to 2023.

Given the imbalanced nature of the dataset, traditional metrics like accuracy, precision, and recall may not accurately reflect model performance. Therefore, this analysis focuses on the F1 score and the area under the precision-recall curve to assess model effectiveness. AUCPR excels in evaluating classifier performance in imbalanced datasets by offering precise insights into the detection of the minority class with low prevalence and significant separation between classes.

3. Result

The GitHub repository contains both raw and cleaned datasets, wherein the cleaned datasets are the outputs of data cleaning and preparation processes and are utilized as the inputs in the machine learning algorithms for the pipe failure prediction. The repository also includes the code used for data cleaning and preparation, enabling researchers to easily execute the provided code to clean and prepare the data, either the open Kitchener dataset or any other dataset.

In scenarios utilizing a random split, the XGBoost algorithm demonstrates superior performance. Conversely, when employing a time-based split for data segregation, the LightGBM algorithm outperforms in effectiveness. The random split method showed impressive results in both training and testing phases, as depicted in Table 1. However, this method's drawback becomes evident when predicting future outcomes, in which its effectiveness significantly declines, indicating that random splits may not yield reliable forecasts. Conversely, while the time-based split method produced consistent results across training and testing phases, it still provided unreliable predictions for the future. These observations underline the importance of evaluating break prediction models against data from future time periods not included in the training dataset.

Table 1. Comparison of model performance for two different strategies of data preparation.

Algorithm	Split	Test AUC-PR	Test F1	Last 5 Years AUC-PR (2018–2023)	Last 5 Years F1 (2018–2023)
LightGBM	80/20 (Randomly in 1985–2017)	0.79	0.78	0.47	0.17
	Interval (1985–2017)	-	-	0.4	0.36

Table 1. *Cont.*

Algorithm	Split	Test AUC-PR	Test F1	Last 5 Years AUC-PR (2018–2023)	Last 5 Years F1 (2018–2023)
XGBoost	80/20 (Randomly in 1985–2017)	0.79	0.78	0.51	0.45
	Interval (1985–2017)	-	-	0.34	0.31

4. Discussion

The establishment of an open repository serves as a template for collaborative model development, evaluation, and benchmarking. It enables direct evaluation and comparison of techniques, empowering researchers and utility stakeholders to make informed decisions regarding the deployment of predictive models. Ultimately, by more accurately forecasting water main breaks, future models can facilitate early interventions and the prioritization of actions, effectively preventing interruptions in the water supply. Continuing to investigate improvements to the model's efficacy, such as by studying other sampling methods and integrating pipe aggregation, is suggested.

Author Contributions: Methodology, F.B.a. and R.D.; validation, F.B.a. and R.D.; writing—original draft preparation, F.B.a.; writing—review and editing, R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Natural Sciences & Engineering Research Council, grant number RGPIN-2022-04664.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are openly available in https://open-kitchenergis.opendata.arcgis.com and https://climate.weather.gc.ca (accessed on 31 March 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kabir, G.; Tesfamariam, S.; Francisque, A.; Sadiq, R. Evaluating risk of water mains failure using a Bayesian belief network model. *Eur. J. Oper. Res.* **2015**, *240*, 220–234. [CrossRef]
2. Barton, N.A.; Farewell, T.S.; Hallett, S.H.; Acland, T.F. Improving pipe failure predictions: Factors effecting pipe failure in drinking water networks. *Water Res.* **2019**, *164*, 114926. [CrossRef] [PubMed]
3. Fan, X.; Wang, X.; Zhang, X.; Yu, X.B. Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors. *Reliab. Eng. Syst. Saf.* **2022**, *219*, 108185. [CrossRef]
4. Snider, B.; Mcbean, E.A. Improving time-to-failure predictions for water distribution systems using gradient boosting algorithm. In Proceedings of the 1st International WDSA/CCWI 2018 Joint Conference, Kingston, ON, Canada, 23–25 July 2018.
5. Chen, T.Y.; Vladeanu, G.; Yazdekhasti, S.; Daly, C.M. Performance evaluation of pipe break machine learning models using datasets from multiple utilities. *J. Infrastruct. Syst.* **2022**, *28*, 05022002. [CrossRef]
6. Omar, A.; Delnaz, A.; Nik-Bakht, M. Comparative analysis of machine learning techniques for predicting water main failures in the City of Kitchener. *J. Infrastruct. Intel. Res.* **2023**, *2*, 100044. [CrossRef]
7. Dziedzic, R. Impact of data preparation on the performance of pipe break status prediction models. In Proceedings of the World Environmental and Water Resources Congress 2023, Henderson, NA, USA, 24 May 2023.
8. Rosenberg, D.E.; Filion, Y.; Teasley, R.; Sandoval-Soils, S.; Hecht, J.; Van Zyl, J.; McMahon, G.; Horsburgh, J.; Kasprzyk, J.; Tarboton, D. The next frontier: Making research more reproducible. *J. Water Resour. Plan. Manag.* **2020**, *146*, 01820002. [CrossRef]
9. Kitchener GeoHub. Available online: https://open-kitchenergis.opendata.arcgis.com (accessed on 31 March 2024).
10. Historical Climate Data—Climate—Environment and Climate Change Canada. Available online: https://climate.weather.gc.ca (accessed on 31 March 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.