1 - Team Lead and Members

Team Members***	Roles and Responsibilities	Skillset
Imran Motala M.Eng., P.Eng., PMP Manager, Water & Wastewater Asset Management Region of Peel Imran.motala@peelregion.ca	• Created a vision for this initiative • Developed project team • Provided guidance and mentorship	• Infrastructure Planning and Asset Management • Hydraulic Modeling and Hydro informatics • Management Consulting and Mentorship
Naysan Saran Co- Founder, CEO CANN Forecast Naysan.saran@cannforecast.com	• Technical project management • Supervise data analysis process and AI model development • Supervise and participate in of model real-time deployments of the data pipelines and operationalization of AI models	• Software architecture design • DevOps • ML and statistical modeling
Nimarta Gill, M.Eng. Advisor, Water and Wastewater Asset Management Region of Peel Nimarta.gill@peelregion.ca	• Lead the project and liaison between the consultant and Region experts • Monitored project progress and evaluated project performance throughout • Accountable for project completion within allocated budget and timelines	• Infrastructure Planning • Asset Management • Capital Budget planning for linear water system • Condition Assessment
Julien Magne Full-stack Developer, CANN Forecast julien@cannforecast.com	• Integrate GIS data of watermains with Excel break history into a format compatible with the AI model • Development of the data quality control, data analysis, and likelihood of failure dashboard	• GIS analysis • Python development • DevOps
Rachel Laplante, Data Analyst CANN Forecast rachel@cannforecast.com	• Data quality control • Data analysis	• Python programming • Exploratory data analysis • Machine learning
Benoit Roland, PhD. Data Scientist, CANN Forecast benoit@cannforecast.com	• Participate in the improvement of the AI model for the detection of most at-risk watermains	• Python programming • Machine learning • Statistical modeling

*** No changes to the team have been made over the course of the project

2 - Problem Statement

Concisely describe the problem/need the Team is addressing

The Regional Municipality of Peel (Peel Region) supplies water to approximately 1.44 million residents and 175,000 businesses across three municipalities in Southern Ontario: the Cities of Mississauga and Brampton, as well as the Town of Caledon. Covering 1,225 square kilometers (473 square miles), the Region is one of the fastest growing areas in North America: its population is expected to increase by 3.7 million inhabitants between 2001 and 2031, which represents approximately 80% of the population growth in the province of Ontario[1].

Thanks to a capital spending of CAN$ 1 billion to replace water mains over the past 20 years, the Region of Peel has reduced the system's length share of its Cast Iron and Ductile Iron pipes from 15% and 25% in 1995, to 1.7% and 8.1% respectively in 2021. In the vast majority of cases, these pipes were replaced by PVC water mains. As a consequence, the Region's break rate fell by 48% from 2010 to 2018 and is still decreasing according to current data. Globally, Peel Region has an excellent risk index with an average annual likelihood of failure of 0.26% per pipe and 2.6 breaks / 100 km over the past five years, which is among the lowest break rates in the world.

In response to population increase and the resulting ever-growing demands on the water supply system, the Region of Peel has implemented a continuous improvement strategy to proactively manage its water infrastructure. As part of this proactive approach, the Region has collaborated with CANN Forecast to determine whether innovative methods such as Artificial Intelligence and Machine Learning could identify the water mains cohorts that are most at risk of failure, allowing decision-makers to predict more accurately the remaining life of the assets and to prioritize renewal programs.

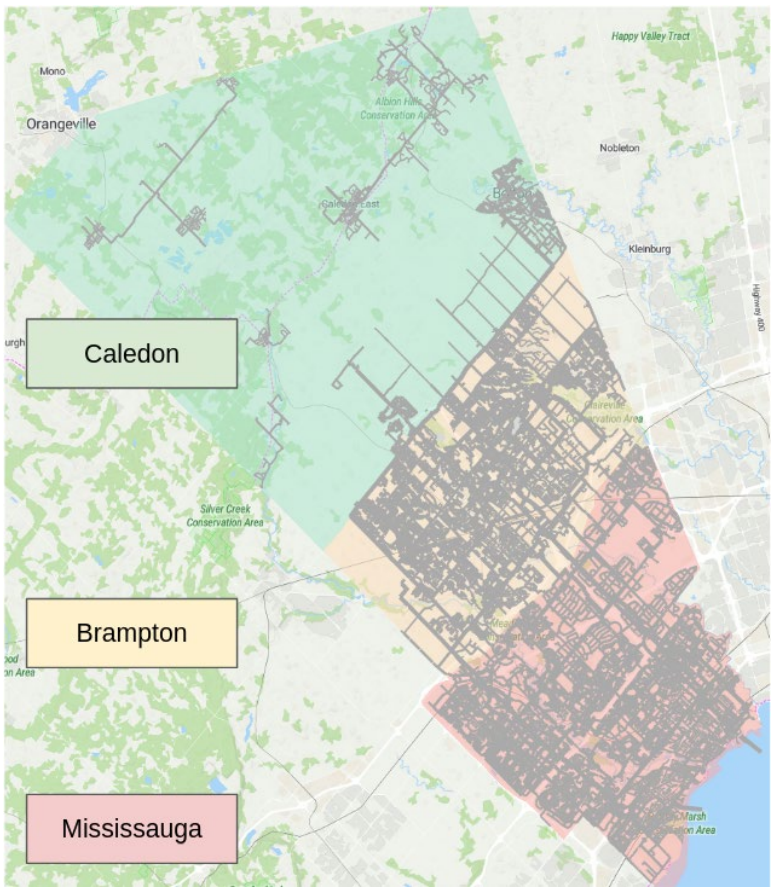

Figure 1: Map of the Peel Region's Water Distribution System

Briefly describe the existing system/conditions (e.g., data source, technology used, networking, system architecture, O&M) that are relevant to the problem being solved

In terms of overall risk-assessment, the Region has worked with a consultant in the past to evaluate the Consequence of Failure (CoF) of its linear water infrastructure. Region also developed water and wastewater Decision Support System (DSS) for linear asset. It is a tool that helps to prioritize the investment decision making to meet level of service approved by Council. It is a SQL based tool that consumes inputs from various sources, runs stored procedure and provides outputs into Excel, GIS files etc. and becomes the starting point for various Asset Management (AM) program development. The use of the tool has increased confidence in short & long-term forecasts to support budget and financial planning processes. Developing a data-driven Likelihood of Failure (LoF) quantification for its watermains was the last component needed to complete the risk analysis. In terms of data sources, the Region has made investments in the past to collect good quality data about its water network and associated break history, as presented below.

Water Network Data

The water network data was provided by Peel Region to CANN Forecast in Esri Shapefile format and contained information regarding the pipes' geometry, diameter, material, length, pressure zone, road class, and soil type, among others. In addition, the "Municipality Name" attribute allowed the distinction between pipes that were installed in Mississauga, Brampton, or Caledon. Finally, data corresponding to abandoned water mains was also provided. For these particular pipes, the abandoned date was set as the date at which they were replaced with a new water main.

Break History Data

Since the year 2000, the Region of Peel has been maintaining extensive records of its water main break history, which include the material, diameter, and the soil type surrounding the pipe that experienced a failure. In addition to these features, break type and break cause were also recorded in approximately 65% of failures. Since the true moment at which a failure occurs can seldom be known with certainty, the work order date was used as the nearest approximation to the actual break date. In general, breaks were linked to their corresponding water main assets using the "From Node" and "To Node" column values while also ensuring that the "Material", "Diameter", and "Installation Date" columns were matching across both data sources. However, when this approach was impossible, the break address value was used to locate the nearest pipe with matching installation date, material, and diameter within a reasonable radius from the break location. Figure 2 shows an overview of the main characteristics of the data that was used for modeling purposes.

[1] https://www.peelregion.ca/pw/water/environ-assess/pdf/2007-MasterPlanReport.pdf

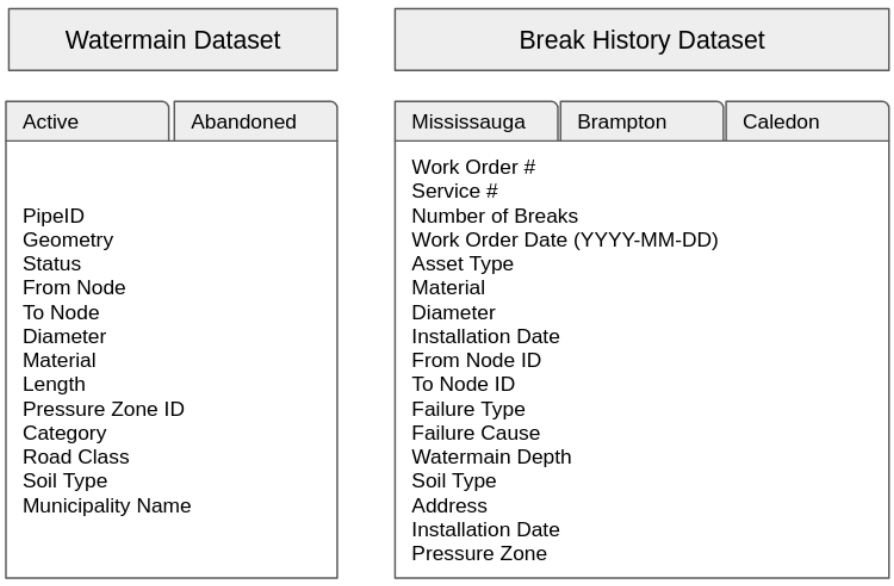

Figure 2: Overview of the water main and break history datasets

Describe key considerations and the desired outcome

The desired outcome of the Challenge Project is to learn whether Machine Learning can be successfully applied to predict future breaks accurately within the water distribution network, thus helping staff from Peel Region to optimize their water main replacement program investments. To be successful, the trained AI model should be able to correctly predict a significant proportion of future breaks for 24 months after its predictions are generated.

The key considerations to measuring the success of the Challenge Project are:

- The Machine Learning model should identify a relatively small proportion of the watermain network (less than 2%) that is responsible for a high proportion of pipe failures. Predictions made by the model should be validated against future breaks to ensure the model's reliability.
- The model should clearly outperform conventional approaches to pipe replacement, namely:
    - replace oldest pipes first
    - replace pipes with most breaks first
- The AI model should not be a black box and should provide staff with reasons to explain its decisions.
- The model should provide insights on the degradation of PVC watermains within the Region's network, as it is now composed of this material at more than 70% and long-term degradation mechanisms of PVC are still generally unknown.

Solution

Describe the proposed solution indicating the value proposition.

The proposed solution aims at employing Machine Learning to predict future breaks in the water distribution system using GIS information about the water network's structural characteristics and past break history. If the AI model proves to be effective at predicting future breaks, its forecasts will be integrated into the Asset Management Decision Support System (DSS) of the Region to optimize future investments.

The proposed value of the solution is to help Peel Region to develop an efficient life-cycle management strategy for its water infrastructure assets and maintain excellent levels of service despite an important population increase over the past few years causing ever-growing demands on the water supply system.

Indicate whether third-party software is required to implement the solution.

- The watermain dataset is owned by the Region and requires a GIS software to be interpreted (ArcGIS or QGIS).
- The machine learning/AI software for pipe failure prediction and associated dashboard was developed by CANN Forecast.
- The Asset Management Decision Support System (DSS) has been developed internally by the Region of Peel.

If a machine learning model is utilized, describe the algorithm and the approach to train, test, and update/retrain the model

The Machine Learning model utilized was developed by CANN Forecast in partnership with McGill University and the Institut National de la Recherche Scientifique (INRS) in Canada. The goal of the model is to identify pipe cohorts that are most likely to fail, where a pipe cohort is defined as a relatively homogenous population of pipes that are expected to have similar physical, environmental, and operational characteristics, and therefore similar degradation curves and performance.

The algorithm leverages unsupervised learning to explore the feature space of watermain and break history, along the following axes: installation date, age, diameter, length, pipe location, soil type, road class and break history. The optimization problem that the model tries to solve is to maximize the difference between the deterioration curves of each cohort and while minimizing the variance within each cohort. Furthermore, all cohorts explored by the system must have a cumulative pipe length of at least 4 Km to ensure statistical reliability in the forecasts.

As a first part of the project, data analysis was performed using water network and break history information between 2000 and 2020. However, because the Peel water network is very dynamic - the Region has reduced its share of Iron watermains from 40% to approximately 10% of the system thanks to a capital spending of CAN$ 1 billion - the years 2015 to 2020 were given a much heavier weight in the training phase, as shown in Figure 3.

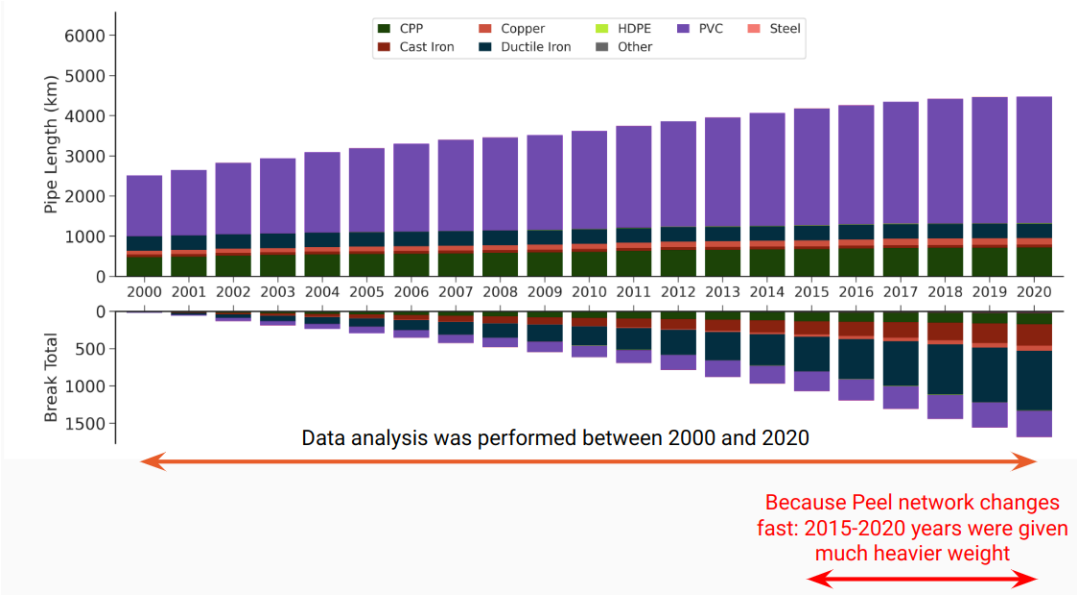

Figure 3: Model train set: years 2015 to 2020 were given a much heavier weight

Indicate hardware/devices utilized for the solution

On the data science pipeline level, the hardware is composed of two Ubuntu 20.4 Linux virtual machines (VMs) hosted on the Azure Cloud. Each of these VMs has at least 8 virtual CPUs and 16GiB RAM. The historical and live data are stored in secured PostGIS databases. A Jenkins pipeline has been created to process new data once it has been pre-validated by a Data Analyst.

Describe application security and architecture

Information is stored on secure cloud servers (Microsoft Azure) and the HTTPS protocol to encrypt all web-based applications. Moreover, within the codebase, access keys are encrypted using git-secret and all of our development servers are secured with the SSH protocol.

The system architecture of the solution is composed of the following components:

- Data quality analysis and quality control (QA/QC) module: flag errors and inconsistencies within the water network and break datasets
- Data formatter: transform the output of the QA/QCA module into a format that can be read by the AI model
- Training module: identification of the most at-risk watermain cohorts
- Dashboarding module: computation of degradation curves and key statistics to be displayed on the dashboard.

Identify data streams and QA/QC consideration

To prevent data errors from creeping into the process and ultimately leading to suboptimal decisions, a preliminary quality control step was performed on the data. Based on best practices and data quality standards from the United States Environmental Protection Agency (EPA, 2002) and Quebec's Center of Expertise in Urban Infrastructure (CERIU, 2014), the quality control system developed by CANN Forecast automatically processed the water network and break history data to identify potential human errors, spelling mistakes, misclassifications and inconsistencies in pipe installation date, diameter, material, and break history. Table 1 provides an overview of the baseline QA/QC rules that were used to flag inconsistencies between pipe material, installation date and diameter.

Material	Acceptable Installation Period	Acceptable Diameter Range
Ductile Iron	1960-present	75-1600 mm
Cast Iron	1850-1970	75-1500 mm
Polyvinyl chloride (PVC)	1970-present	100-1200 mm
Steel	1850-present	100-3600 mm
High-density polyethylene (HDPE)	1968-present	100-1600 mm
Concrete Pressure Pipe (CPP)	1900-present	350-3600 mm
Asbestos Cement	1900-1980	75-1050 mm
Copper	1950-present	12-300 mm

Table 1: QA/QC rules applied to pipe material, installation date and diameter

As a result of this initial step, 95% of pipe segments and 93% of breaks were declared anomaly-free, confirming that they could be directly used for modeling purposes. The remaining data were flagged according to the type of error found and were either discarded or fixed in collaboration with staff from the Region of Peel.

Describe any difficulties faced during the development/deployment of the solution and how the team mitigated them

The following difficulties were faced during the development and deployment of the solution.

Difficulty	Details	Mitigation
Integration of the type of soil into the model	The soil information within the watermains GIS data was divided into the following types: clay, clay loam, loam, variable, sand, organic and silt. The breaks dataset also contained soil type information that was considered more accurate as these recordings were made by crews in the field during repair. However, the breaks dataset contained 27 different types of soil that could not readily be matched with the watermains soil types. Furthermore, the number of mismatches were too numerous to be reconciled.	The watermains soil type categories were used because of their ease of integration into the algorithm.
Combination of sub-networks	One conceptual difficulty about this project was related to the fact that the Region of Peel is composed of three distinct municipalities: Mississauga, Brampton and Caledon. At the beginning of the project, it was unclear whether we should develop one model per municipality, or use the whole network composed of all three cities as the training set for the AI algorithm.	In the end, it was decided to train one model on each municipality's network separately and compare the decision trees generated for each sub-network in terms of similarities and differences. In retrospect, this approach was very useful as we realized that the algorithm had generated very similar decision trees for Mississauga and Brampton PVC pipe cohorts, even though the training sets were separated, while the Cast and Ductile Iron cohorts were much more at-risk in Mississauga compared to the other municipalities.
Time required to validate the model's performance	One of the main objectives of this project was to ensure that made by the model would be validated against future breaks to ensure the model's reliability. This meant that both teams (Peel and CANN Forecast) were not satisfied by the model's performance on historical data *in hindsight*, but actually had to wait 24 months for enough new watermain breaks to accumulate before conclusions could be drawn regarding the model's success in predicting pipe failures.	The training and phase of the project was performed in 2021 on historical watermain and break data from 2000 to 2020 inclusively. The validation phase was conducted in 2023 using new breaks from 2021 to 2022.

Provide key performance indicators to quantify the performance/benefits of the solution. Compare the performance of the system using the proposed solution vs. status quo/conventional approaches.

To ensure that the AI model was indeed able to predict future breaks, this project spanned three years:

- During the first year (2021), the algorithms were trained on historical watermain and break data from 2000 to 2020 inclusively and generated a list of high-risk watermain cohorts to be monitored:
    - The highest-risk cohort was Ductile Iron with three breaks or more, and shorter than 381 m in length. This cohort totaled 5.4 Km of linear watermains as of 2021
    - The second highest-risk cohort was also made of Ductile Iron, this time with two breaks, and shorter than 362 m in length. This cohort totaled 5.1 Km of linear assets as of 2021
    - Finally, the third most at-risk cohort was made of 7.5 Km Cast Iron

- In total, these three most at-risk cohorts represented a total of 15 miles, which is less than 1% of the total water network of the Region

- Between 2021 and 2022, breaks were recorded for the validation phase of the project

- In 2023, the validation phase was performed. The goal following key performance indicators were calculated:
    - KPI1: What proportion of the 2021 and 2022 breaks were correctly predicted by the 15 miles most-at-risk watermains?
    - KPI2: How did this performance compare to conventional approaches to pipe replacement, namely:
        - replace oldest pipes first
        - replace pipes with most breaks first

KPI1: What proportion of the 2021 and 2022 breaks were correctly predicted by the 15 miles most-at-risk watermains?

As shown in Figure 4, the top 3 most at-risk cohorts - identified by the model using 2015-2020 data - were responsible for 14% of 2021 breaks and 22% of 2022 breaks. This is an excellent overall performance as they represent less than one percent of the network, which typically is close to the total linear length that most utilities can replace or reline within two years.

Because the validation was made on future data over two years, this KPI suggests that by using the algorithm, the Region has the potential of reducing its yearly breaks by a significant percentage, by focusing on the most at-risk pipe cohorts, which represent less than one percent of its water distribution infrastructure.

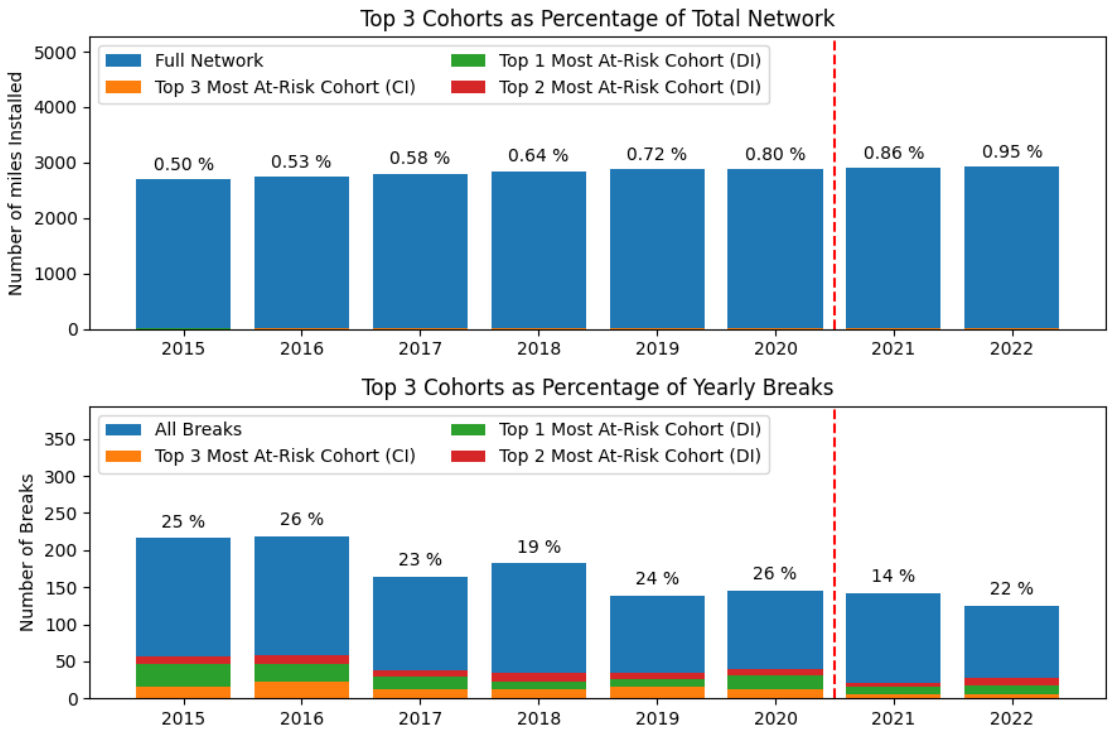

Figure 4: The top 3 most at-risk cohorts were responsible for 14% of 2021 breaks and 22% of 2022 breaks

Figure six presents the spatial distribution of the 2021 and 2022 breaks for the Region. This figure also shows that the cohorts' performance trends are coherent despite clear year-to-year variations in the break spatial distribution pattern.

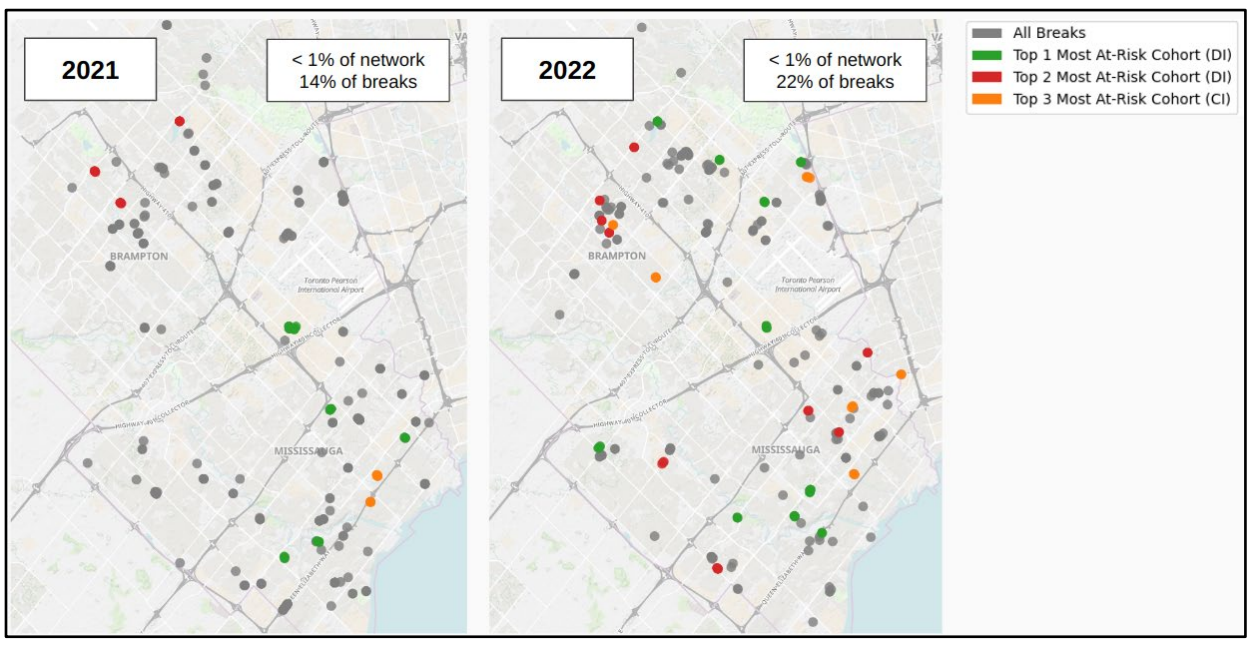

Figure 5: Coherent good model performance despite the variation in the year-to-year break pattern

KPI2: How did this performance compare to conventional approaches to pipe replacement?

Since the top 3 most at-risk cohorts identified by the algorithm had a total of 15 miles, their performance was compared with the following two conventional approaches to pipe replacement:

- Replace the oldest 15 miles of watermains within the Region's network
- Replace the 15 miles of watermains with the highest number of previous breaks

As shown in Figure 6, the model clearly outperforms these conventional approaches to pipe replacement:

- In 2021:
  - 50% increase compared to replacing watermains with higher number of previous breaks
  - 2000% increase compared to replacing the oldest pipes first
- In 2022:
  - 211% increase compared to replacing watermains with higher number of previous breaks
  - 833% increase compared to replacing the oldest pipes first

It is worth noting that for utilities that do not have access to AI-based tools, replacing pipes with the highest number of previous breaks is much more efficient than replacing their oldest watermains in general.

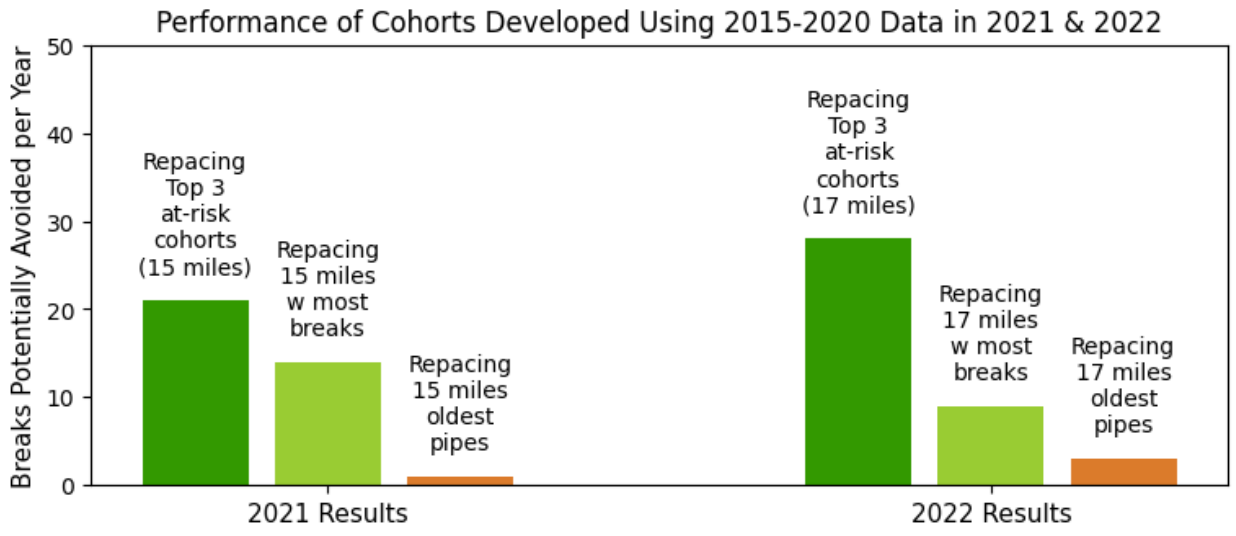

Figure 6: Top 3 most at-risk cohorts performance compared to conventional approaches

Describe whether the solution has been implemented. If not, describe how the solution would be deployed/implemented, including implementation cost, infrastructure requirements, schedule, etc.

The solution has been implemented in Region's State of Good Repair watermain replacement budget. For year 1 i.e., 2024 budget, only the highest-risk cohort were integrated into the DSS and based on COF score were recommended for inspection program to get information regarding the structural condition of the pipe.

Region has collaborated with CANN Forecast to update the analysis based on recent watermain break data. The new run will enable the Region to identify new high-risk cohorts of mains based on the condition assessment data and water repair work orders. The most at-risk cohorts will be prioritized in next 1-3 years budget which will allow Region to maximize the return on investment with regards to the replacement program.

Describe the approach to scaling the solution to larger systems/systems with more data.

As part of this project, the small diameter distribution watermains were considered ranging from 50mm-500mm in diameter. The intent is to automate the process by using the break data for next few years at least to identify the trend of breaks and incorporate more data such as the effect of pressure/ transient in the metallic mains along with the impact of live and dead load on a local versus collector road.

For this project, only the drinking water network was analyzed. The solution could be scaled to integrate wastewater and road infrastructure as well.

Indicate how/if the results of the implementation will be communicated and used by the utility.

The GIS deliverable provided by CANN Forecast's team has already been added to Region's Decision Support System (DSS) as an additional layer so that analysis of pipe's structural, hydraulic, and operational performance can be done in comparison to if the pipe has been identified to be part of a high-risk cohort using AI learning. The co-relation of this information gives Region a clear understanding of the degradation curves of watermains of different material.

Next Steps for the solution beyond the Challenge

The future steps for this solution beyond the Challenge include:

- Integration of road traffic loads and cathodic protection as new inputs to the AI model
- Continuous improvement of the cohorts by integrating the latest repair, rehabilitation, replacement and pipe installation data