Predicting Watermain Breaks in the City of Ottawa

Niki Jafari (101224861), Yujing Yang (101216350), Mahitha Sangem (101212458)
Supervisor: Prof. Olga Baysal

I. ABSTRACT

This project proposal outlines the analysis of variables including the time of the watermain breaks, pressure zones, the diameter and material of the watermains, the frost depth, the temperature in Ottawa, and their role in predicting watermain breaks. Data on watermain breaks in Ottawa will be used to identify which variables contributed to these breaks. This research will help the City of Ottawa determine where condition assessments will need to occur.

II. KEYWORDS

Watermain, Watermain Breaks, Condition Assessment, Condition Assessments, Frost Depth, Pressure Zone, Temperature, Time Analysis, Watermain Diameter, Pipe Material, Predictive Modelling, Machine Learning, Statistical Analysis, Regression Models.

III. INTRODUCTION

Watermains refer to the pipes that distribute water throughout Ottawa. The purpose of the Ottawa Water Distribution project is to predict watermain breaks. This project is important because in recent years, watermain breaks in Ottawa have led to sinkholes, flooding, traffic, and road closures.

The water system in Ottawa is large and essential to citizens' lives. There are hundreds of people working at the branch daily. Employees are focused on maintenance throughout the winter and focused on construction throughout the summer. Therefore, analyzing watermain breaks will assist with the decision-making process in terms of allocating funds to the appropriate watermains, and predicting rehabilitation and replacement.

Predicting watermain breaks through this analysis will enable the City of Ottawa to allocate funds to maintain, rehabilitate, and replace at-risk pipes. The City of Ottawa can also use cameras and other tools to closely monitor at-risk pipes and avoid major failures through their condition assessment program in the summer.

The research questions are (1) What factors contribute to watermain breaks in the City of Ottawa? and (2) How can we predict watermain breaks?

IV. RELATED WORKS

A Canadian case study analyzed 45 years of watermain break records (more than double any reported literature, including 86,000 pipe breaks and 11,000 km of buried watermains) for five major water utilities, to determine long-term break trends [1]. All of the Canadian utilities analyzed either experienced consistent break rates or a significant decrease in break rates in the past 10 years [1].

Despite this, Canadian surveys consistently state aging infrastructure as the most significant challenge facing utilities, and several studies indicated that pipe break rates have increased [1]. 28% of pipes in service in North America today are over 50 years of age and are approaching their expected end-of-service, and watermain breaks have increased by 27% in the last six years [3].

Therefore, watermain replacement is necessary to prevent watermain breaks and subsequent water contamination and service interruptions. Nonetheless, this is an expensive process, and utilities must find ways to prioritize which watermains to rehabilitate or replace. To address this, pipe break prediction models have been

developed to assist utilities in prioritizing pipe replacement, to reduce future pipe breaks [2].

There are numerous studies examining the best methods to predict watermain breaks. The most popular methods include simplistic models (relying on a single factor to rank the likelihood of watermain breaks, such as age), physical models (infield measurements to determine the stressors acting on a pipe), and machine learning and other statistical models [2]. While simplistic models often dismiss several important factors that may contribute to watermain breaks, and physical models are time-consuming and expensive, machine learning and statistical models are an effective and more affordable way to predict watermain breaks [2].

V. ANTICIPATED RESULTS

Prior to the analysis, it was predicted that (1) The relationship between the pipe material and pipe diameter are highly correlated with watermain breaks, (2) There is a certain time of day in which watermain breaks occur more frequently, likely in the mornings, and (3) A frost model will predict the frost depth based on the temperature change.

VI. METHODOLOGY

Tableau is used for data visualization. This includes the overall trend of watermain breaks from 2001 to 2020, pressure zone break counts, as well as the relationship between pipe material and diameter with the number of watermain breaks.

The statistical analysis includes regressions, and Chi-square hypothesis tests focusing on (1) Pipe Material, (2) Pipe Diameter, and (3) the Time of the Watermain Break.

Machine learning models are used to forecast the frost depth based on temperature, as severe cold is also likely to influence the occurrence of watermain breaks.

VII. DATASET DESCRIPTION

The analysis was conducted using two datasets provided by the City of Ottawa.

Dataset 1 was over the span of approximately 20 years, with data on watermain repairs. This data consisted of variables including pipe material and diameter, pressure zones, and the time and location of the breaks. This data will enable the identification of pipes that will need monitoring or preventative repairs.

Dataset 2 consisted of the frost depth and temperatures in Ottawa for the months of November to March, over the span of approximately 10 years.

VIII. IMPLEMENTATION

A) Statistical Analysis

i) Historical watermain break record

The historical watermain break record contains 20 years of data in Ottawa from 2001 to 2020. By conducting data visualization using Tableau, an overall trend can be seen. As evident in Fig. 1, there is a slightly downward facing trend. It shows that overall, watermain breaks are decreasing throughout the years.

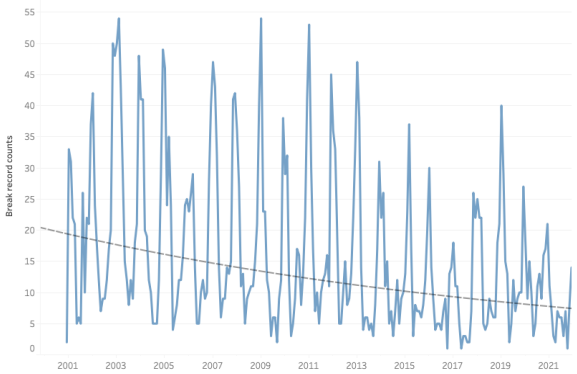

Fig. 1. Watermain Break Record 2001 - 2020

ii) Pipe Material and Diameter analysis

By looking at the data, it is evident that the relationship between pipe material and diameter with watermain breaks is worth examining. Since the relationship between two independent factors and one dependent factor is being examined, a stacked bar graph is used. Pipe material was placed on the x-axis, the count of watermain breaks on the y-axis, and the pipe diameter as a different stack.

As evident in Fig. 2, three pipe materials have the most break counts: CI, DI, and UCI. Also, the 152mm, 203mm, and 305mm diameters are the most frequently broken pipes. A hypothesis has been formulated: The smaller the pipe diameter, the more likely it is to break. However, further analysis is required to test this hypothesis.

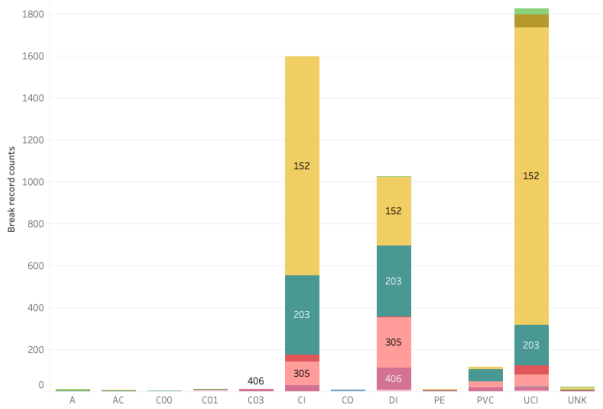

Fig. 2. Pipe Materials and Diameter Counts

To test the hypothesis, a regression was performed using R. An alpha of 0.05 was used in this regression.

The results of the simple regression analysis showed that the p-value of the independent variable was greater than 0.05, and the association between watermain break frequency and pipe thickness was not significant.

It is evident that pipe diameter is not necessarily correlated with the frequency of watermain breaks.

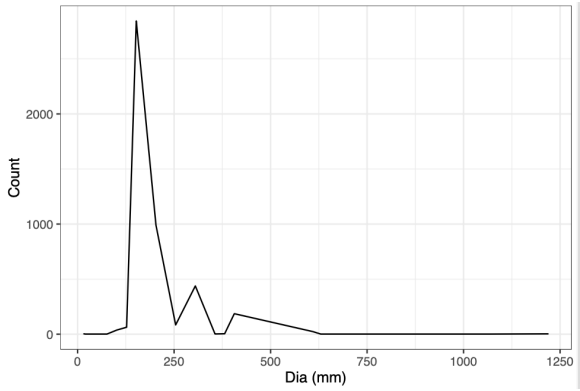

Fig. 3. Watermain Break Count Based on Diameter

iii) Pipe pressure zone analysis

By looking at the data, it seemed that the relationship between pipe pressure zone and watermain break frequency was worth analyzing.

A box and whiskers plot was chosen because there is a range of watermain break counts for most of the pressure zones. The box and whiskers plot for watermain breaks and pressure zones shows that 2W2C, 1E, and 1W are the pressure zone codes in which watermain breaks were most likely to occur.

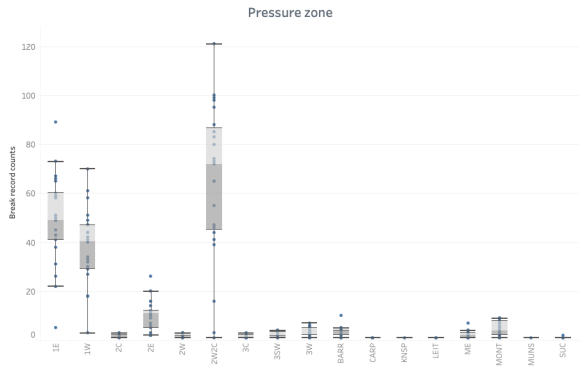

Fig. 4. Pressure Zone and Break Record Counts

iv) Time analysis

Another factor to consider is to determine if there is a certain timeframe in which watermain breaks occur more frequently. First, four hours as a timeframe were used to examine if there is a significant difference between different

timeframes and the watermain break record counts.

The simple bar graph performed using R showed that the timeframe of 8:00 to 12:00 had a significantly larger number of watermain breaks. A statistical hypothesis method needed to be used to test these results.

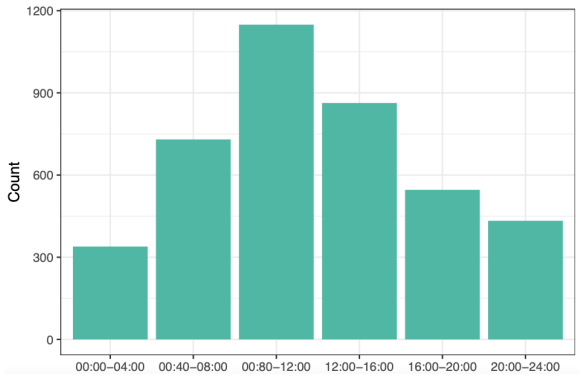

Fig. 5. Watermain Break Count Based on Time for every four hours

A Chi-square test was chosen because it is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if there is a difference between observed data due to chance, or if it is due to a relationship between the variables. In this case, the two variables are the time of the watermain breaks and the count of the watermain breaks.

In order to use a Chi-square test, an eight-hour timeframe was chosen for the analysis. This is because a four-hour timeframe was too small as the result would be obvious.

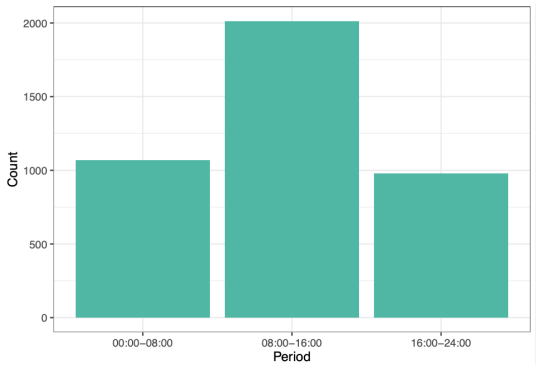

Fig. 6. Watermain Break Count Based on Time for every eight hours

With this result, a Chi-square test was used to test whether the period 08:00 to 16:00 was when watermain breaks occurred most frequently, compared to other times.

In the time period of 08:00-16:00, the frequency of pipe breaks was the highest. A Chi-square test was performed on the time period, and the result showed a p-value < 2e-16. This shows that there was a significant difference in watermain breaks between the different timeframes, and the 08:00-16:00 time period had significantly more breaks than other time periods. This may be due to the increased frequency of use.

B) Frost Depth Prediction

Climatic changes also affect watermains, in which frost has a major impact. This occurs when the temperature goes below a certain threshold such as 0 degrees, which results in the freezing of water in the soil. The lesser the temperature, the deeper the frost formation will be. This adds pressure underground and increases the downward force on the pipes, which makes the underground pipes vulnerable to fracture or break.

The second dataset consists of data on the temperature and frost depth (in cms and inches) in Ottawa over the past 10 years during winters, specifically November 2011 to January 2022. The dataset had each year's data separate, which were then concatenated in order to perform prediction, which added up to 1600 records.

i) Analysis

The temperature features are Mean Temperature and Cumulative Mean Temperature, as the variation in these factors impacts the frost depth, and hence are considered as features for predicting frost depth. Data framing and a thorough analysis was conducted on these features, which would have an impact on the depth of frost formation. Fig. 7 shows the histogram and scatter plots of Mean Temperature and Cumulative Mean Temperature, to show this data aggregated and as a collection of points.

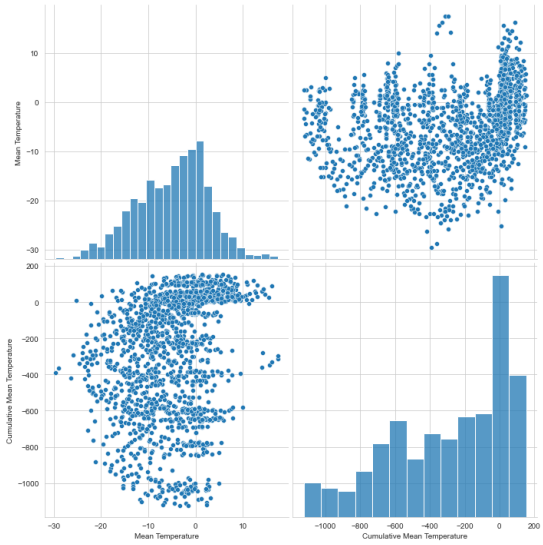

Fig. 7. Histogram and Scatter Plots of Mean Temperature and Cumulative Mean Temperature

Fig. 8 shows the distribution plots of Mean Temperature and Cumulative Mean Temperature, where plotting of data distribution of these features against the density was performed.

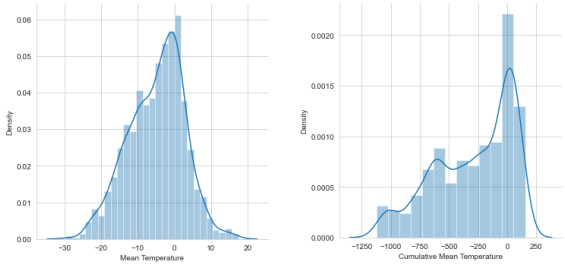

Fig. 8. Distplots of Mean Temperature and Cumulative Mean Temperature

Fig. 9 shows the Probability Density Function (PDF) and Cumulative Distribution Function (CDF) plots of Mean Temperature and Cumulative Mean Temperature. PDF is a graph that counts the number of failures between different time periods, resulting in a curve that estimates the number of failures you can expect at a particular number of time units. The CDF plot depicts the calculations of the likelihood that a random observation from the population will be less than or equal to a certain value.

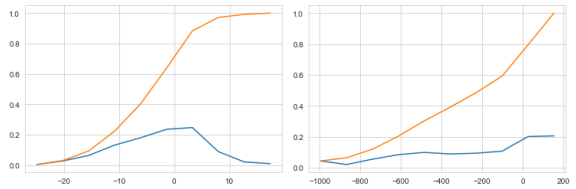

Fig. 9. Probability Density Function (PDF) and Cumulative Distribution Function (CDF) plots of Mean Temperature and Cumulative Mean Temperature

Fig. 10 shows the mean, standard deviation, median, quartiles, percentiles and mean absolute deviation of the features Mean Temperature and Cumulative Mean Temperature.

```
Mean:
-4.917584480600753
-292.0638297872338

Std-dev:
7.594406901977302
-292.0638297872338


Median:
-3.8499999999999996
-220.29999999999998

Quantile:
[-10.1   -3.85   0.4   17.4 ]
[-577.9 -220.3   10.8  152.8]

90th Percentile:
3.7
71.50999999999996

Median Absolute Deviation:
7.635401425303849
371.0953352919521
```

Fig. 10. Mean, standard deviation, median, quantiles, percentiles and mean absolute deviation of the features Mean Temperature and Cumulative Mean Temperature

Fig. 11 shows the box plots of the features Mean Temperature and Cumulative Mean Temperature, to display the distribution of data based on a five-number summary (a “minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).

Fig. 11. Box plots of the features Mean Temperature and Cumulative Mean Temperature

Fig. 12 shows the violin plots of the features Mean Temperature and Cumulative Mean Temperature, in addition to box plots they also show the probability density of the data at different values.

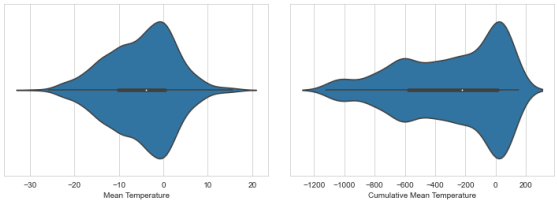

Fig. 12. Violin plots of the features Mean Temperature and Cumulative Mean Temperature

Fig. 13 shows that there is a minimum correlation between the selected features i.e., Mean Temperature and Cumulative Mean Temperature, which is a requirement of the features, in order to proceed with the prediction.

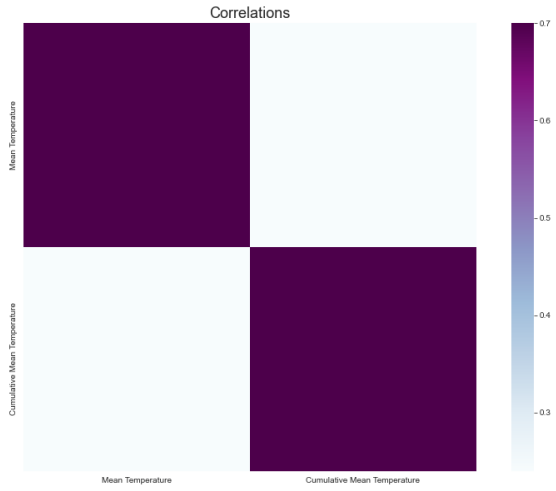

Fig. 13. Correlation between Mean Temperature and Cumulative Mean Temperature

ii) Prediction

Since the prediction to be conducted is in the form of continuous values, different regression models were used to perform the prediction. The 1600 records of data were split between training and testing datasets: November 2011 to March 2018 for training, and November 2018 to January 2022 for testing, making it a split of around 70% and 30%, respectively. Scaling was implemented for these values to enable the models to be trained, such as learning and understanding the pattern being followed. The training data was fed to the regression models and their predictions were tested using the testing data.

The regression models used for the prediction are Linear Regression, KNN Regression, SVM Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression. Fig. 14 shows the performance of the first model i.e., Linear Regression, with the prediction plot in blue and actual values in red. It also portrays the green prediction line and a dashed line for ground truth, which is the actual values' plot. Similarly, Fig. 15 shows the performance of a KNN Regressor. Fig. 16 shows the prediction of an SVM Regressor. Fig. 17 shows the prediction of a Decision Tree Regressor. Fig. 18 shows the prediction of a Random Forest Regressor. Fig. 19 shows the prediction of a Gradient Boosting Regressor.

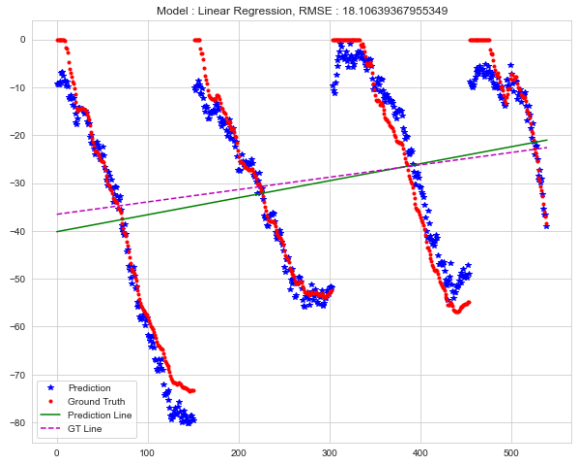

Fig. 14. Linear Regression Prediction

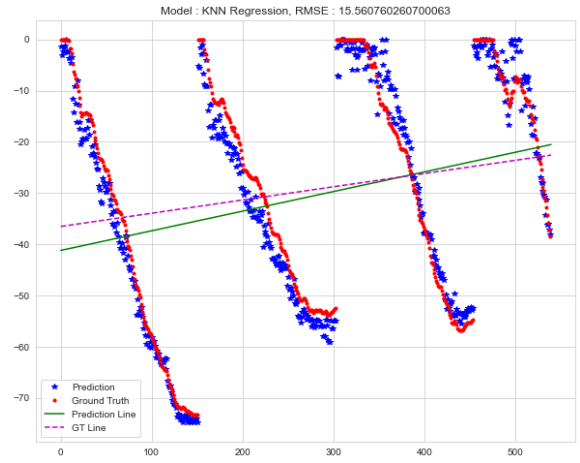

Fig. 15. KNN Regression Prediction

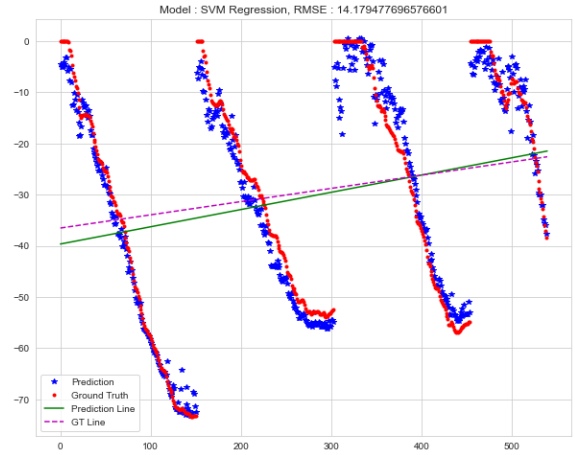

Fig. 16. SVM Regression Prediction

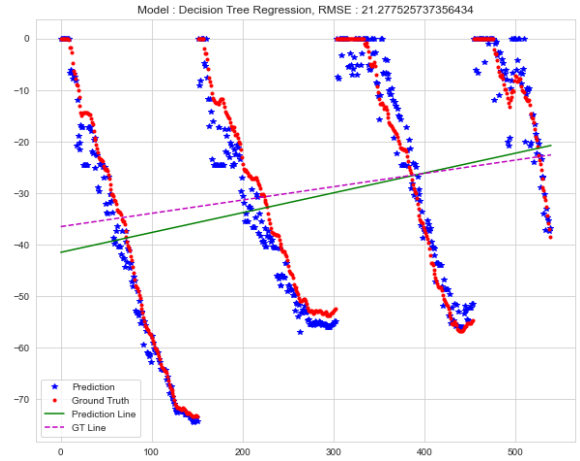

Fig. 17. Decision Tree Prediction

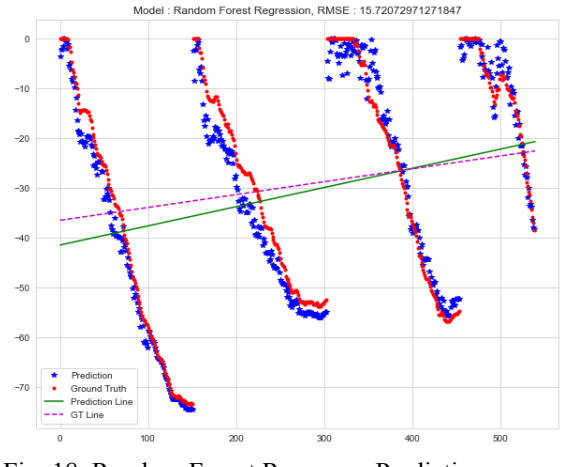

Fig. 18. Random Forest Regressor Prediction

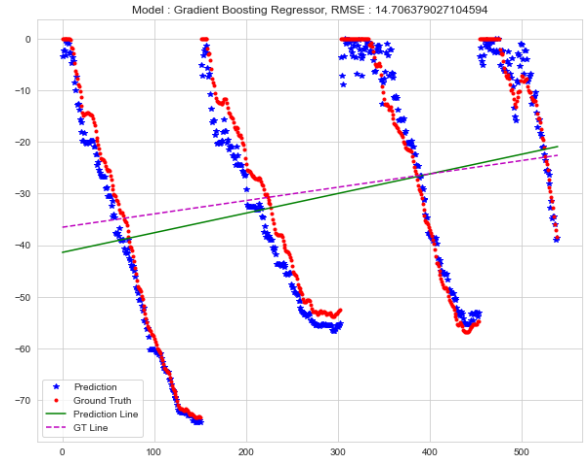

Fig. 19. Gradient Boosting Regressor Prediction

The performance of these models was analyzed using the performance metrics: R2 score, Mean Absolute Error, and Root Mean Square Error (RMSE), as shown in the tables below Fig. 20 and Fig. 21. These metrics determine how well the regression model predictions fit the actual/observed data.

Regressor	R2-score
Linear Regression	0.96
KNN Regression	0.97
SVM Regression	0.97
Decision Tree Regression	0.96
Random Forest Regression	0.97
GradientBoostingRegressor	0.97

Fig. 20. R2 score of models

Regressor	Mean Absolute Error	RMSE
Linear Regression	3.35	18.11
KNN Regression	3.22	15.56
SVM Regression	3.02	14.18
Decision Tree Regression	3.45	21.67
Random Forest Regression	3.15	15.97
GradientBoostingRegressor	3.13	14.71

Fig. 21. Mean Absolute Error, and Root Mean Square Error (RMSE) of models

From the metrics, it was observed that the SVM Regressor gives the best performance, i.e., Maximum R2 score, and minimum mean-absolute-error and RMSE.

IX. RESULTS

There are several results based on the implantation:

i) The overall watermain break situation in Ottawa is improving throughout the years.

ii) Three pipe materials have the highest break record counts: CI, DI, and UCI.

iii) The association between break frequency and pipe thickness was not significant.

iv) 2W2C, 1E, and 1W are the most frequently broken pressure zone codes.

v) There is a significant difference between the time periods of pipe breaks, and 08:00-16:00 has significantly more watermain breaks than other time periods.

vi) The frost prediction implementation shows that, for the frost depth, the SVM Regressor provides optimal prediction.

X. LIMITATIONS

One of the limitations of this research is that the data lacked variables such as the age of the watermains, the depth of their installation, and the temperatures of the soil and water. Moreover, there was no data on the watermains that did not break. This information could have contributed to a more accurate prediction of watermain breaks in Ottawa. With additional variables, including data on watermains that did not break, models can be trained to predict the occurrence of watermain breaks in the future.

XI. CONCLUSION

This research is essential because it is associated with Ottawa citizens' quality of life. When a watermain break occurs, there is potential for significant loss of water. The total cost of water loss due to watermain or pipe breaks is estimated to be 3.8 billion USD per year in North America [2]. Watermain breaks also have serious implications for the citizens of Ottawa, including the contamination of clean drinking water. Other issues include service interruptions, environmental costs associated with the loss of water, and lessening the quality of the water.

Moreover, it is estimated that 500 billion USD is required to replace the aging infrastructure over 25 years [2]. Watermain break prediction models can assist in identifying which pipes need to be replaced in order to reduce future breaks and avoid replacing pipes that are still in good condition. Therefore, the cost savings associated with using pipe break prediction models can be significant (Snider & McBean, 2020).

For future research, collecting additional variables on Ottawa's existing watermains, such as their age will be beneficial. Moreover, analyzing data on watermains that did not break, in addition to existing data on watermain breaks, will enable a more accurate prediction of watermain breaks in the future. To more accurately understand the long-term trends in pipe conditions and break rates, improved analysis of longitudinal datasets containing large numbers of break record years is required [1].

Moreover, as climate change will likely lead to more unpredictable weather and more significant temperature changes, it would be interesting to explore its effects on watermain breaks and predictive modelling.

XII. CONTRIBUTIONS

Niki Jafari was responsible for the background research, writing, editing, and formatting.

Yujing Yang was responsible for the statistical analysis of the factors contributing to watermain breaks.

Mahitha Sangem was responsible for the frost model prediction and analysis.

This research would not have been possible without the guidance of Professor Olga Baysal and the City of Ottawa Water Distribution Team's data and ambitions.

XIII. REFERENCES

[1] B. Snider and E. A. McBean, "State of watermain infrastructure: a Canadian case study using historic pipe break datasets," *Canadian Journal of Civil Engineering*, vol. 48, no. 10, pp. 1266–1273, Oct. 2020.

[2] B. Snider and E. A. McBean, "Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions," *Urban Water Journal*, vol. 17, no. 2, pp. 163–176, Feb. 2020.

[3] S. Folkman, "Water Main Break Rates in the USA and Canada: A Comprehensive Study," *Mechanical and Aerospace Engineering Faculty Publications,* p. 1 March. 2018