Identification of the Most Important Factors Driving Watermain Failure

Sadaf Gharaati

A Thesis
in
The Department
of
Building, Civil, and Environmental Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Applied Science (Civil Engineering) at
Concordia University
Montreal, Quebec, Canada

March 2022

© Sadaf Gharaati, 2022

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Sadaf Gharaati

Entitled: Identification of the Most Important Factors Driving Watermain Failure

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Civil Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________ Chair
Dr. M. Nik-Bakht

______________________ Examiner
Dr. O. Moselhi

______________________ Examiner
Dr. M. Nik-Bakht

______________________ Thesis Supervisor(s)
Dr. R. Dziedzic

______________________ Thesis Supervisor(s)

Approved by ______________________
Dr. M. Nik-Bakht Chair of Department or Graduate Program Director

____

Dr. Mourad Debbabi Dean of Faculty

Abstract

Identification of the Most Important Factors Driving Watermain Failure

Sadaf Gharaati

As essential infrastructure, water distribution systems provide water to the vital needs of individuals, businesses, and industries. Watermain failure jeopardizes water systems' ability to deliver clean water safely. The main target of this study was to identify the most influential factors on watermain failure across Canada. Dimensionality reduction approaches were applied to watermain data from thirteen Canadian cities, Barrie, Calgary, Region of Durham, Halifax, Kitchener, Region of Markham, Region of Waterloo, Saskatoon, St. John’s, Vancouver, Victoria, Waterloo, and Winnipeg. While previous studies have focused on small datasets of a few cities at a time, the present study compares various factors in different networks with different characteristics. Multiple physical, historical, protection, operational, and environmental factors were compared. Two target attributes were defined, current rate of failure and break status. A correlation analysis was applied to each city to identify the relationships between different attributes and the targets. Four dimensionality reduction approaches were employed to evaluate the impacts of different factors on the targets and identify the most important factors The four approaches are Factor Analysis of Mixed Data (FAMD), Categorical PCA (CATPCA), Random Forest Recursive Feature Elimination (RF-RFECV), and Extreme Gradient Boosting Recursive Feature Elimination (XGBOOST-RFECV). Results indicate CATPCA is more reliable than other approaches. Furthermore, protection activities were found to be more important than physical and historical attributes in most utilities. Thus, the collection of protection data should be prioritized for utilities with higher rates of protection activities, especially if they have already collected data on fundamental physical and historical attributes. While few utilities collect data on environmental, operational, and certain physical factors such as roughness, dead-end, restrained, and pipe depth, these were also found to be important and should be further investigated. These findings create the foundation for a new data collection framework for predicting main breaks.

Acknowledgments

I would like to render my special gratitude to Professor Rebecca Dziedzic, my beloved supervisor, for her dedicated support and guidance during my master's study. Completing this research and writing this thesis could not have been possible without her expertise and guidance. Her precious knowledge, patience, and valuable advice helped me walk through this challenging but fascinating journey.

Second, I would like to acknowledge Concordia University for funding the project and technical support during the Covid-19 pandemic.

I wish to express my warmest gratitude to the National Water and Wastewater Benchmarking Initiatives (NWWBI) representatives for providing this study's dataset and valuable feedback during workshops.

Last but not least, I must thank my parents, who encouraged me unconditionally, my brother, who advised me during different stages of this journey, and my fiancé, Mehrzad, for his endless support during this adventure. This achievement could never have been conceivable without him.

This thesis shall be devoted to my fiancé, Mehrzad.

Table of Contents

List of Tables

List of Figures

1 Introduction

Water main deterioration is a global challenge that can jeopardize water systems' ability to deliver clean water safely. The failure of water mains can affect individuals, businesses, industries, and institutions. In order to realize the importance of clean water, imagine a day without it. Water main breaks can directly disrupt the service provided by pipes. According to the United States Environmental Protection Agency (U. S. Environmental Protection Agency, 2012), failure costs can considerably be higher than maintenance costs based on transmission size. According to the Canadian Infrastructure Report Card (2012), the cost of upgrade and replacement of water and wastewater network in Canada is estimated to be more than CAD$ 80 billion. The total cost of water loss resulting from watermain breaks is expected to be 3.8 USD billion per year (U. S. Environmental Protection Agency, 2010). Hence, it is essential for water utilities to seek cost-effective rehabilitation and renewal strategies (Kleiner & Rajani, 2001). These strategies should reduce failure costs while ensuring an adequate level of service.

1.1 The components of a water distribution system

Water supply systems consist of transmission, distribution, and service phases. In the first phase, water from the source, i.e. river, lake, or groundwater, is transmitted to a treatment plant to provide clean water for customers. Once water is treated and leaves the plant, the distribution phase begins. Clean water is transferred to storage and delivered to customers in different areas. The distribution network consists of three major parts of pipes, valves, and flush hydrants.

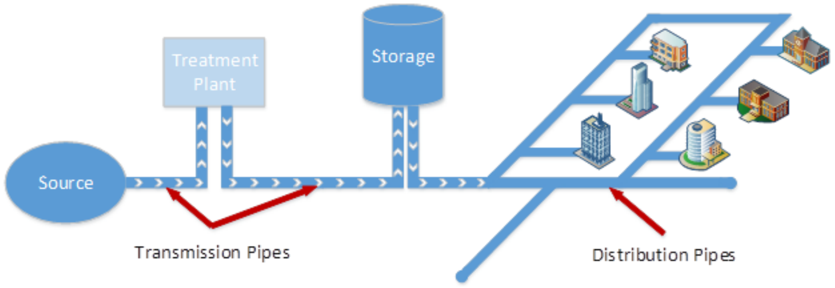


Figure 1-1 Water distribution system (Adapted from EPA, 2006)

1.1.1 Pipes

Pipes convey clean water within the water distribution network. The amount of carried water and the required pressure for the end-user determine the size of the pipes.

Transmission pipes usually carry water from considerable distances, such as the treatment planet, to storage tanks. These pipes are generally the largest in the system (>350 mm diameter) and, accordingly, are the most expensive ones (Karimian, 2016).

Distribution pipes provide water services to all potential users and usually are laid in the city street. Compared with transmission pipes, they are smaller in diameter (<350 mm diameter) (Karimian, 2016). Service pipes transfer water from distribution mains to the end-users ‘buildings or property. Although these pipes are usually small in diameter, Saskatchewan Environment determines the minimum acceptable size of 25 mm (Saskatchewan Water Security Agency, 2004) .

1.1.2 Valves

Valves in the distribution network are used to control flow and isolate a particular part of the system during repair and maintenance.

1.1.3 Hydrants

Flush hydrants in water distribution systems are responsible for removing silt, sediment, rust, etc., from the waterline. These hydrants are usually located at the end of dead-end lines. The flush hydrants also check closed valves and weak flows in the system. Although fire hydrants are larger and more expensive than flush hydrants, some municipalities use fire hydrants instead of a flush hydrant.

1.2 Overall condition of pipes

According to the Water Research Foundation (WaterRF, 2017), 75% of water utilities consider pipe breaks the key pipe replacement measure. The Canadian Infrastructure Report Card (CIRC, 2019) indicates that about 30% of potable water infrastructure in Canada is in very good condition, 40% in good condition, and 25% in fair, poor, or very poor condition. Canadian Infrastructure Report Card (CIRC, 2019) briefly summarizes these conditions as Table 1-1.

Table 1-1 Condition rating definition adapted from CIRC (2019)

Condition	Description
Very good	The asset is suitable for the future. It is maintained well, is in good condition, and might be new or recently rehabilitated
Good	The asset is in adequate condition and generally is at the middle stage of its expected service of life. The condition of this asset is acceptable
Fair	Signs of deterioration and deficiencies in some elements have appeared. These assets require attention.
Poor	The condition can highly affect the provided service by the asset. The majority parts of assets have deteriorated significantly, the expected service of life is approached, and the condition is below the standard.
Very poor	The expected service of life is near or passed and demonstrates the sign of advanced deterioration. The asset is not suitable for sustained service. Some of the assets might be unusable.

Approximately 30-40% of these assets are aged 20 years. A water pipe's typical expected service life is between 70 and 100 years, depending on the material and local characteristics. Pipes can be categorized into three major material groups: plastic, concrete, and steel. Plastics generally have a higher ESL and are currently used more frequently.

According to a survey of 308 utilities in the USA and Canada, the break rate grew by 27% from 2012 to 2018 from 11 to 14 breaks in 100 miles each year in which significant increase in break rate of asbestos cement and cast iron pipes is a cause of concern for water utilities (Folkman, 2018). In general, PVC pipes have the lowest break rate. Respondents also indicated the average length of pipe replaced is 0.8% of the total length each year. Moreover, the length requiring replacement doubled from 2012 to 2018 (Folkman, 2018).

1.3 Problem statement

Various factors are contributing to watermain failure and previous studies have focused on a different subset of data. The studies also have focused on one or two cities at a time and results have not been broadly implemented. Developing a reliable prediction model requires comparing a comprehensive range of factors that affect pipe failure. Also, the contributing factors might be highly correlated and including them, all in the analysis can lead to incorrect results. Additionally, collecting all the factors for utilities is usually costly and time-consuming. A comprehensive study comparing various factors in different area is required to develop a reliable data collection framework.

1.4 Objectives

The main objective of this study is to identify the most important factors in predicting watermain failure. In order to achieve this target, the following sub-objectives are taken into consideration:

1. Compare the applicability of the different dimensionality reduction approaches across 13 cities in Canada;

2. Identify the most important factors affecting watermain current rate of failure in each of the cities based on the results of dimensionality reduction approaches;

3. Identify the most important factors in predicting if a pipe will experience a failure in each of the cities based on the results of dimensionality reduction approaches; and

4. Develop a data collection framework based on the final results.

2 Literature Review

2.1 Factors contributing to water main failures

Pipes usually follow a typical life cycle represented by a “bathtub curve”. The curve represents three life cycle stages: burn-in, in-usage, and wear-out. These stages describe the period after installation with failures due to defective pipes or installation problems, a trouble-free period, and a period of increasing breakage rate due to aging and deterioration, respectively (Kleiner & Rajani, 2001).

Factors contributing to water main failures can be categorized into general physical factors, historical information, protection activities, environmental factors, manufacturing defects, and operational factors. Barton et.al., (2019) reviewed different factors in previous watermain deterioration studies and discussed the following typical factors influencing breakage:

- Material
- Pipe joint system
- Coating and lining
- Manufacturing process
- Handling, storing, and third-party damages
- Age
- Diameter
- Environmental factors such as cold/warm temperature, seasonality, and soil movements
- Operational factors such as internal water pressure
- Previous failures

These factors determine the material's applicability to different pressure conditions and soil types.

2.1.1 Physical factors

2.1.1.1Material

Pipe material can be categorized into four general groups: Metallic, Plastic, Concrete, and Asbestos cement. Among them, Cast Iron (CI), Ductile iron (DI), Steel, Asbestos Cement, polyvinyl chloride (PVC), and polyethylene (PE) are most common. Different materials have unique break patterns and expected service lives (Snider & McBean, 2020). Results of a survey for 308 utilities in Canada and the USA indicate that circumferential breaks and corrosion are the two most common types of break in DI and CI pipes (Folkman, 2018). For PVC pipes, scratches, void, and inclusion are most common (USEPA, 2012). A study of Regina's water distribution system indicated that more than 90% of the AC pipes that failed between 1994 – 2003 had a circumferential break (Hu & Hubble, 2005).

The expected service life (ESL) of an asset defines the years during which adequate service is provided (Hudson & Haas, 2013). Although the true ESL depends on various factors, such as design and construction methods, usage and environmental condition, maintenance, and operation practices (Hudson & Haas, 2013), high-level ESLs have been defined by pipe material type.

For example, the USACE (1998) recorded ESL of concrete pipes between 70-100 years and around 50 years for steel and plastic pipes. A study of the City of Cobalt, ON (2014) considered 100 years for ESL of all pipes, including PVC, Copper, and DI., while according to the Municipal Association of South Carolina (MASC), metallic pipes, including DI and CI, have an ESL between 100-120 years and plastic pipes such as PVC and HDPE are expected to last 70 years (2016).

Over the years, improvements in pipe manufacturing processes of different materials have led to changing trends in material use. A study by Kirmeyer (1994) estimated in 1992 that DI and CI covered two-third of watermains in the USA, AC pipes around 15%, and the remaining either plastic or concrete. However, during recent years this pattern is changed. According to a survey result of 308 utilities in the US and Canada, DI and PVC are currently the most common pipe materials. (Folkman, 2018). According to the same study, in USA and Canada, no breaks were recorded for AC and CI pipes after the 1980s, which indicates these pipes were not been widely used since then. Also, DI and PVC pipes failures had not appeared before 1960, meaning that these materials began to be applied later. A study by Snider and McBean (2020) on five large utilities across Canada also confirmed this pattern.

Plastic pipes, in general, are the only type of material resilient to corrosion. Although AC is more corrosion resistant than metallic materials, it is rigid, and signs of corrosion can be observed in AC pipes buried in sulfate soils or close to acidic ground water. Previous studies indicate that the vulnerability of metallic material against corrosion is different. For instance, Ductile iron is the least vulnerable metallic material to corrosion, followed by steel and Cast iron, respectively (Hou et.al., 2016) (Barton et.al., 2019). Since steel is cheaper than DI and is suitable for higher pressure, it is recommended for large diameter pipes (>800 mm). At the same time, DI, which is more resilient to corrosion and has higher tensile strength, is suggested for pipes with a diameter between 300 and 800 mm (Barton et.al., 2019). A study of large diameters pipes (>300mm) in five Australian utilities by Rajeev et.al. (2015) indicated that among cast iron, ductile iron, steel, and AC, the lowest break rate is for Ductile iron. In contrast, the highest break rate belongs to unlined and cement-lined Cast iron. The failure rate of DI pipes is considerably more than CI pipes (Snider & McBean, 2020). According to the same study, this can be rooted in the younger age of these pipes and more flexibility of them due to the existence of graphic nodules which can protect the pipe against ground movements.

2.1.1.2Diameter

While pipes with smaller diameters (<200mm) are frequently used in water systems, their failure rate is the highest, as compared to larger diameters (Barton et.al, 2019). Around 67% of installed pipes in US and Canada have a diameter of 200 mm or less (Folkman, 2018). Results of a study by Hu and Hubble in Calgary (2007) indicate that more than 93% of AC pipe breaks in the city were of 150-200 mm pipes. This can be related to factors such as the low resilience of these pipes against ground movements, their thickness, corrosion, and poor joint reliability. Diameter can also affect failure mode. For instance, the Circumferential break type is common in small diameter pipes (<200mm) (Bruaset & Sægrov, 2018) while, longitudinal and hole breaks are primarily observed in large diameter pipes (>300mm) where water pressure is generally higher (Rajeev et.al., 2015).

2.1.1.3Length

Pipe length is identified as an important factor by many authors. However, there isn't a consensus about whether breakage is positively or negatively correlated with breaks. For instance, results of a study by Berardi et.al. (2008) on deterioration of a UK water distribution system suggested pipe length as one of the three most important factors in explaining deterioration. It also confirmed length linearly and positively affects pipes burst. On the other hand, a negative correlation between pipe length and the number of breaks is concluded from other studies. A study of watermains in Quebec City, QC, by Wang et.al., (2009) indicates a lower annual break rate was recorded for longer pipes with the same age and diameter. Nishiyama and Filion (2014) also reached this same conclusion, who found breaks to be highly negatively correlated with length in a study of cast iron pipes in Kingstone.

2.1.1.4Depth

Pipe depth refers to the depth where a pipe is buried, measured from the top of the pipe. Previous studies have indicated that increasing buried depth results in higher soil temperatures during the winter (Rajani et.al., 1996), thereby reducing the probability of failure at higher depths. A frost line is defined as the maximum depth in which ground water in the soil is expected to freeze. Increasing the depth after this point may no longer reduce breaks. Environmental factors such as weather conditions, soil characteristics such as soil type and heat transfer properties, and nearby heating sources are considered by cities to determine freezing depth. For instance, while in Vancouver's mild climate, pipes are buried at a depth of 0.6m for frost protection (Brown, 2006), in Toronto, they are buried at 1.8m depth where they are believed to be just below the frost line (City of Toronto, 2020). This depth is even greater for Winnipeg, with harsher winters, around 2.4m (Manitoba Soils and Screw Piles, 2018).

2.1.1.5Thickness

Pipe wall-thickness negatively correlates with failure rate, i.e., thinner pipes have higher break frequency. This is observed in a study of ductile iron pipes in Sanandaj, Iran (Asnaashari, McBean, Shahrour, & Gharabaghi, 2009). Another study by Bruaset and Sægrov (2018) related lower corrosion resilience of small diameter pipes to their thinner wall thickness. In general, increasing pipe wall thickness increases pipe resilience to circumferential loads and prevents pipe wall reduction due to corrosion and other degradation (Rezaei et.al., 2015). Also, pipe failure due to high pressure is more likely when the pipe wall thickness is reduced. A study by Snider and McBean (2020) indicates that increasing wall thickness of the pit cast manufacturing process installed before 1920 can increase ESL of CI pipes compared with thin wall spun cast iron installed between 1920-1940.

2.1.1.6Joint type

Joint failure is a typical pipe failure mode (Barton et al., 2019). Results of studies by Dingus, et al., (2002) and Burn, et al., (2005) reported joint failure was the cause in average, 15% and 16% of all PVC pipes failure. Joints connect pipe segments and are either built as a part of the pipe, i.e., integral, or installed separately at two pipe ends. According to Rahman and Farrell (2007), integral joints are cheaper to be installed, less prone to corrosion, and less prone to hum-related errors, which leads to a lower probability of leakage. Joints can also be categorized as rigid or flexible according to their ability of residency against rotation and displacement. Examples of rigid joints include mechanical bolted joints and flanged, whereas flexible joints include socket joints and spigots. Since rigid joints are sensitive to rotation, they are more prone to leak and fracture failure due to ground movement (Barton et.al., 2019). In general, corrosion and leakage are the most common cause of rigid joint failure, while gasket failure is the main cause of flexible joints (Barton et.al., 2019).

2.1.2 Historical information

2.1.2.1Age

Age is one of the most important breaks predicting factors (Berardi et.al, 2008) since it can indicate many other factors such as corrosion, external loads, deterioration, etc. Also, age is used in various statistical models such as time exponential and time powered to describe the time dependency of breakage and to estimate an optimal time for pipe replacement. It can be said that the failure rate increases as the pipe age increases. However, age alone cannot be a reliable factor in predicting watermain (Kahn et.al., 2020). According to the bathtub curve, the failure rate can be high immediately after pipe installation. However, this rate drops for a certain period and then increases as the age increases (Kleiner & Rajani, 2001).

Various studies observe increasing break rates for older pipes. Physical damages, environmental and operational conditions which a pipe might experience during its life are all factors which contribute to the growth of break rates over time.

2.1.2.2Previous break

It is also observed that, pipes are more likely to break once they have broken before (Kleiner & Rajani, 2001). This fact is addressed by several researchers such as Walski and Pelliccia (1982) , Mark et.al, (1985), Andreou et.al., (1987a) (1987b) related data are reflected in many prediction models. Previous breaks can provide a proxy for local conditions such as area soil type, weather condition, usual traffic load, etc. Accordingly, previous break information can be a more informative factor than age. Snider and McBean (2020) analyzed five Canadian utilities that showed varying break year trends depending on various factors, notably weather. Accordingly, it is essential to analyze break rates over many years to find a general pattern. The same study results indicate a generally constant or decreasing long-term break rate.

2.1.3 Protection activities

2.1.3.1Lining and Coating

Lining and coating are pipe protection methods performed inside and outside of the pipes, respectively, in order to delay the corrosion process. For instance, lining a pipe that is in reasonably good structural condition can extend ESL by 30-50 years (USEPA, 2002). The structural lining is expected to last long-term, around 50 years (U. S. Environmental Protection Agency, 2002). According to the same study, these linings improve pipe strength against dynamic loads and can act as a pipe for a short period when the pipe has failed.

The most common types of coating used are resin, PE sleeve, and yellow jacket, whereas bituminous, cement mortar, and resin are more applied as linings. According to Guan (2001), the most common type of lining in North America is bituminous.

Resin corrosion protection consists of adding a layer of resin to a pipe. This layer hardens over time and creates a non-penetrable layer that prevents pipes from being exposed to surrounding water (Wiley, 2018).

Yellow jacket coating consists of two layers of polyethylene which enable long-lasting and resistance against corrosion, and most biological, chemical, and environmental contaminants find in the soil.

PE sleeve isolates pipe from soil and creates a uniform layer of passive water which prevents corrosion (Malizio, 1986). It is believed that although initial rusting might occur, oxygen in trapped water will be consumed gradually by cathodic reactions and prevent further corrosion (Rajani & Kleiner, 2003).

Cement mortar creates a layer of passive iron oxide. This layer is then kept in an alkaline environment of hydrated Portland cement (Bardakjian, McReynolds, & Hausmann, 2007). This lining material is economical and capable of withstanding various operational conditions.

Bituminous hydrophobic property repels water and moisture from vulnerable parts. This product's durability, flexibility, and chemical attack resistibility enable it to be widely used in harsh environments (Nanan, 2019).

2.1.4 Environmental factors

2.1.4.1Soil and ground characteristics

Soil characteristics such as PH, humidity, and soil type can also affect pipe failure. The influence of soil on pipe failure can be direct, i.e., soil movement, or indirect, i.e., corrosion. In the following paragraphs, soil-related factors are discussed.

2.1.4.1.1 Corrosion

Soil is an essential factor affecting corrosion. Risk of corrosion by soil is defined as the risk of soil-related electrochemical attack or chemical actions which result in corrosion in un- protected pipe material. Lower PH and resistivity and a higher level of moisture in the soil are factors that facilitate corrosion. A corrosion index can measure soil corrosivity. This metric indicates corrosion potential and can either relate to soil corrosivity or water corrosivity. Various factors are included in different studies for calculating the soil corrosion index. Among them, the American Water Works Association (AWWA) defines a numerical scale for corrosion index and rates soil corrosivity based on soil characteristics, including resistivity, Redox Potential, sulfide content, and moisture. These scores range between 0 and 10, with higher indicating higher corrosivity.

The effects of corrosion can be observed in terms of wall- thickness reduction and functionality interruptions in pipes. Numerous failure modes are also corrosion-related (Folkman, 2018). Folkman (2018) reported 75% of 308 surveyed utilities in the USA and Canada are experiencing at least one area with corrosive soil condition and indicates corrosion is responsible for 28% of watermain failure in the area which after circumferential breaks is the highest. A study by Hu and Hubble (2007) suggests that chemical attacks in Calgary dramatically affected AC pipe's structural integrity. The most common corrosion protection approaches are anode installation, cathodic protection, lining, and coating. A study by Snider and McBean (2020) confirms plateauing and decreasing of cast iron pipes break rate in the 1990s when the city began a cathodic protection rehabilitation program. Before that, the failure rate of cast iron pipes was increasing. This pattern is also valid for other metallic material pipes such as Ductile Iron.

2.1.4.2Weather

Weather conditions are mainly related to sudden seasonal changes, temperature, frost, and precipitation. It is believed that weather condition plays a vital role in pipe failure. The increasing failure rate of AC and PVC pipes during dry summer and metallic pipes in winter have confirmed this. Frost heave, i.e., ground swelling due to freezing temperatures, results in additional distribution network loads and damages (Bruaset & Sægrov, 2018). Multiple freezes and thaw cycles further increase pipe failures (Bruaset & Sægrov, 2018). Colder temperatures, in particular, increase circumferential breaks, and the effects of extreme temperatures on smaller diameter pipes are greater (Rajani et.al., 1996). These factors affect soil and stability and can lead to pipe breakage. According to previous observations, a lower failure rate is observed during the spring when soil is wet, and ground movement is unlikely (Barton et.al, 2019).

Increasing water consumption and lowering ground water levels impose additional internal and external loads on pipes during dry and hot summers. The internal loads occur due to increasing pressure and velocity in and the pipes, whereas the additional external loads result from soil shrinkage (Wols & Thienen, 2014). Moreover, when soil temperature increases, pipes expand longitudinally during hot weather. However, soil and pipe water temperature differences result in further thermal pipe stress. Finally, higher temperatures can also accelerate corrosion (Wols & Thienen, 2014).

2.1.5 Manufacturing defects

Some early breaks can be rooted in manufacturing defects. The type of defect varies depending on material type. For instance, for metallic pipes, the most common are non-uniform wall thickness, cold shots, and micro cracks. For AC pipes, uneven distribution of asbestos fibers is the primary type of defect. Inclusions and porosity lead to early failure in both metallic and Plastic pipes.

Inclusion is the addition of unwanted material to the pipe, leading to inconsistencies in the material texture and weak points. These defections affect the structure of pipes and can lead to eventual failure. Porosity defection is caused by trapped air in the solidified melted iron mold, leading to cracks in pipes.

Other damages on pipes can be related to storage, specifically for PVC pipes, since long exposure of this material to ultraviolet light embrittles it (Barton et.al., 2019).

According to Snider and McBean (2020), Canadian cities without historical pipe information should predict breaks for CI pipes installed after 1941. On average, 50% of pipe length exceeds AWWA's break rate threshold of 0.125 brks/km/yr.

2.1.6 Operational factors

2.1.6.1Pressure and External Load

Sudden changes in internal pressure can add additional stress on pipes and increase the risk of failure. Pressure-related failures are more common in large diameter pipes since smaller diameters are generally used in lower pressure areas.

External loads result from external factors such as traffic, frost, and soil. The ability of pipes to handle these loads depends partly on the support from surrounding soil, so-called bedding. For instance, for a flexible pipe material, i.e., PVC, lack of proper bedding will result in vertical deflection, leading to joint leakage. However, in rigid pipes, improper bedding will directly lead to pipe breakage since lack of flexibility in these pipes limits pipe reaction to loads.

2.2 Break Type

Break type refers to the form of failure. Different factors can lead to different kinds of breaks. Thus, other models can be developed depending on the type of failure. Since maximum pressure on pipes depends on pipe size and material, these two factors usually determine break type. Breaks can be categorized into circumferential, longitudinal, split bell, and hole groups. Circumferential breaks rooted in longitudinal stress on pipes. Either temperature-related contraction on pipes causes the longitudinal stresses, or soil movements and improper bedding such as insufficient bedding practices and large voids in the bedding near pipes related.

Longitudinal breaks stem from excessive pipe pressure or external loadings such as traffic and frost. The frost penetration leads to the expansion of frozen moisture in the ground and greater loads. The mentioned factors result in transverse stress and contribute to longitudinal break. If this break occurs at the joint, then it is Also the transverse stress affecting the pipe joints that causes split bell breakages.

The other failure mode is related to holes caused by corrosion. Corrosion explained in detail in section 2.1.4.1.1.

As mentioned earlier, Circumferential breaks are the most common failure mode of small diameter pipes while longitudinal and hole are widely observed in larger diameter pipes.

2.3 Pipe deterioration models

In order to prioritize water main replacement and rehabilitation, prediction models are essential (Kleiner & Rajani, 2001). Previous studies have predicted water main breaks through various physical and statistical models. These predictions can reduce operating costs, service level impacts, and health risks of customers. Either approach can provide deterministic or probabilistic results.

2.3.1 Physical models

Physical-based methods focus on the structural performance of water mains and consider internal and external loads. External loads can be defined as the loads transmitted to a pipe from the external surroundings; frost, traffic, and soil overburden are perfect demonstrations of such loads. On the other hand, internal loads in water networks are systems of forces caused by operational pressure and affect the inner side of pipes.

The following are some examples of physical-based methods.

Rajani and Zhan (1996) and Zhan and Rajani (1997) introduced methods for calculating frost load on buried pipes in trenches and under roadways, respectively. Results suggested that using a material with an equal or lower frost susceptibility than the side wall can reduce the development of frost load.

Kuraoka et.al., (1996) provided information on the effects of changes in internal and external pressures and temperature changes on jointed water main. Their analysis showed that the reduction of pipe size might increase maximum axial stress in ductile iron and PVC water main.

The need to measure internal and external loads generally means physical methods are complex and require long data collection and analysis periods and higher costs. These models are usually applied for significant projects where failure costs are high. The present study mainly focuses on statistical methods applied to various types and levels of data.

2.3.2 Statistical models

Statistical-based approaches use historical data on pipe breakage for identifying breakage patterns and assume these patterns continue in the future.

Various statistical models include time-linear, time-exponential, proportional hazard, Poisson, Markov chain, artificial neural network, rate of failure, and other regressions such as logistic regression, evolutionary polynomial regression, etc., introduced. A brief description of some statistical models are provided in the following paragraphs.

2.3.2.1 Poisson regressions

Poisson regressions are generalized linear models which show what independent variables have the most significant impact on the target variable. Generalized linear models are linear regressions applicable to variables with error distributions other than normal. A Poisson regression assumes a Poisson distribution for the target variable. This regression works only with numerical, continuous variables (Zeileis, 2008).

2.3.2.2Markov Chains

Markov Chain is a probability theory assumes one can make future predictions based on current states (Karsten, 2010). The Markov Chain process consists of a number of states which represent possible scenarios of the system (Grinstead & Snell, 2012). In the case of water mains, these states may represent the number of breaks (Gustafson & Clancy 1999), pipe condition (Hong, 1998), etc.

The process starts in one of the states and proceeds from one state to another successively. Each of these transitions is known as a step. The probability of moving from one state to another is only dependent upon the system's current state.

2.3.2.3Artificial Neural Networks (ANN)

Artificial Neural Networks mimic the human brain and seek to find patterns between inputs and outputs. Input, hidden, and output layers are three essential parts of this model. Data enters the process through input layers, hidden layers are responsible for information processing, and finally, results are produced in the output layer (Najjar & Basheer, 1996). The most challenging part of the process is finding the number of hidden nodes and layers (Asnaashari et.al., 2013).

2.3.2.4Proportional hazard models

Proportional hazard models are regression models which are commonly used for finding the relationship between survival time and different predictors. These models apply a vector of hazards, i.e. a combination of factors that affect pipe survival. In other words, the goal of this model is to examine the effect of various factors on the rate of failure, the hazard rate. In the case of reference, a model developed by Andreou et al. (1987a) and (1987b) and Marks et al. (1985) can predict watermain failures.

2.3.2.5 Time exponential models

Time exponential models, according to Belk (2015), are well-known methods for dynamic systems in which the target variable is a function of time. A function is said to have exponential growth when its derivative with respect to time is a constant number multiplied by the function itself. This expression can be shown as equation (1):

$$\frac{dy}{dt} = ky \tag{1}$$

The solution for this equation for a positive constant value k has the form as function below:

$$y = y_0 e^{kt} \qquad (2)$$

Where:

$y,$ is a time exponential function;

y_0, the initial value of y;

t, time; and

k, growth constant;

2.3.2.6 Time linear models

Time linear models assume the function varies linearly with time. This means the function derivative with respect to time is a constant number, called growth rate**.** Time linear models for water main break prediction define a number of breaks as a linear function of time and other variables. The simplest form was proposed by Kettler and Goulter, where the total number of breaks per year in a pipe is a linear function of its age.

2.3.2.7Logistic regression models

Logistic regression models predict the probability of a binary dependent variable. Imagine a binary target variable Y is modeled as a conditional probability with respect to the function of X, Pr(Y=1|X=x). To do so, rather than defining P as a linear function of X, the logistic regression considers log $(\frac{p}{1-p})$ as a linear function of X. In the case of main breaks, it can be applied to predict whether or not a specific pipe will break during a certain time frame. According to Yamijala et.al., (2009), many utilities would rather have information on the probability of a pipe experiencing at least one break as opposed to the number of breaks since one break is often enough to trigger costly repairs.

2.3.2.8Evolutionary polynomial regression (EPR)

Evolutionary polynomial regression combines genetic algorithms (GA) and least squares (LS) to provide a pseudo-polynomial regression model. GAs mimic the process of natural selection. The first step of the algorithm includes the selection of the fittest individuals from a population. These individuals then produce offspring, which can transfer the characteristics of their parents to the next generation. The fitter the selected individual, the better the child will be and the higher their survival chances. This iterative process continues until a generation with the fittest individual is found (Mallawaarachchi, 2017). The least-squares method is applied to evaluate the fit by minimizing squared errors. Improving the EPR method led to introducing a new concept, multi-objective EPR, in which single objective genetic algorithms are replaced with multi-objective genetic algorithms. This approach increases the accuracy of models by reducing the number of polynomial coefficients and reducing the number of data (Keramati et al., 2014). In a multi-objective EPR, the aim is to provide multiple objectives to control different aspects of the model such as fit, complexity, and physical logic (Romanova et.al., 2014).

Table 2-1 compares different watermain prediction models. As presented in the table, previous studies have employed multiple statistical and machine learning algorithms in watermain prediction models. These models have focused on a few cities at a time and have studied a different subset of data. In the provided table, the machine learning methods are a general group representing various machine learning algorithms such as decision tree, random forest, support vector machine (SVM), naïve bias, gradient boosting, etc. Each study employed a different subset of data and compared different machine learning algorithms. For instance, while, Giraldo-González and Rodríguez (2020) compared four machine learning algorithms, including ANN, naïve bias, gradient boosting, and SVM using multiple physical, historical, and operational factors to predict probability failure in Bogota, Colombia; another study by Aslani et al., (2021) have applied random forest, boosted regression tree, Multivariate adaptive regression splines, and ANN and employed diameter, material, length, age, and the number of failures to predict watermain rate of failure in Tampa, Florida.

Table 2-1 Comparing different watermain prediction models and the considered factors

Model		Time Exponential				Time Linear				Proportianal hazard					Markovian		ANN					Poisson			Mchine Learning							
Author		Shamir and Howard	Walski and Pelliccia	Clark et al	Y. Kleiner & Rajani	McMullen	Kettler and Goulter	Jacobs and Karney	Yamijala et al.	Marks et al.	Jefry	Andreou et al.	Bremond	Vanrenterghem-Raven et al.	Kleiner	Micevski et al.	Moselhi and Shehab-Eldeen	Ahn et al.	Al-Barqawi and Zayed	Tran et al.	Snider and McBean	Miller	Asnaashari et al.	Giraldo-González and Rodríguez	Winkler et al.	Shi et al	Shirzad and Safari	Almheiri et al.	Giraldo-González and Rodríguez	Robles-Velasco et al.	Aslani et al.	Snider and McBean
Year		1979	1982	1982	2004	1982	1985	1994	2009	1985	1985	1987	1997	2003	2001	2002	2000	2005	2006	2007	2020	1993	2009	2020	2018	2018	2019	2020	2020	2020	2021	2021
Location		N.A	Binghamt	2 cities in North America	Ottawa – Carlton	Des Moines	Manitoba	Winnipeg	Texas	New Haven	New Haven	New Haven, Connecticut Cincinnati	Bordeaux	New York	N.A	Newcastle	Montreal	Seoul	Moncton London Longueiul	Greater dandenong	A canadian network	Melbourne	Sanandaj	Bogota	A city in Australia	Totonto	Mahabad- Mashhad	Quebec city	Bogota	Seville	Tampa	A canadian network
Physical	Diameter	■	■	■			■		■				■	■		■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■
	Length						■	■	■	■	■	■		■				■	■		■		■	■	■		■	■	■	■	■	■
	Material	■	■				■		■					■		■	■	■	■	■					■			■		■	■	
	Pipe Depth																		■	■							■					
	Thickness																		■				■			■						
	Condition														■	■																
	Slope																			■												
Historical	Age	■	■	■		■		■		■	■	■		■									■	■	■		■		■	■	■	
	Vintage			■						■	■	■																				
	Installation year						■										■	■	■	■	■											
	Number of Breaks	■	■	■	■		■				■	■					■	■	■		■								■	■	■	■
	Time between breaks												■																			
	Break Type	■	■	■			■											■	■				■									
Protection	Lining status																				■											■
	Cathodic protection				■																											■
Operational	Service type																								■							
	Pressure									■	■	■										■	■		■		■			■		■
	Overhead traffic													■								■										
	Cover																						■									
	Number of hydrants																												■			
	Number of valves																												■			
Environmental	Soil Type	■	■				■		■					■		■		■	■		■	■										■
	soil PH					■																			■							
	soil humidity								■																■				■			
	Soil resistivity					■																			■							
	Sulfide																								■							
	Ground condition												■																			
	Land use								■	■	■	■																	■			
	Precipitation								■									■											■			
	Temperature								■									■														
	Water quality																	■														
	Number of trees																			■												
	Road rating																															
	Location																			■												■

2.4 Dimensionality reduction

Since various factors can affect pipe breakage and previous research has focused on different subsets of data, finding the most effective available factors is essential.

Reducing the number of variables is formally referred to as dimensionality reduction. There are two main approaches for dimensionality reduction, feature elimination, and feature extraction. The first reduces the number of variables by eliminating some, whereas the latter creates new independent variables from combinations of previous independent variables. In feature extraction, since new variables are combinations of old variables, none of the original data is eliminated.

2.4.1 Principal Component Analysis

One well-known method of feature extraction is Principal Component Analysis (PCA). PCA is considered as the heart of dimensionality reduction methods. In it, a set of p original variables can be replaced by an optimal q number of derived variables which may provide the ability to represent the full data set. The quality of a PC can be evaluated based on the proportion of the variance that the PC carries (Jolliffe & Cadima, 2016). PCA seeks to find patterns in the data by identifying differences and similarities through eigenvectors and eigenvalues of covariance matrices. The following paragraphs describe the definitions of covariance matrices, eigenvectors, and eigenvalues, which are crucial to PCA.

2.4.1.1The covariance matrix

A covariance matrix is a symmetric square matrix in which each element on its principal diagonal represents the variance and off-diagonal elements, covariance. Since the correlation matrix is the covariance matrix divided by standard deviation, it can be said that a covariance matrix is a non-normalized form of the correlation matrix.

2.4.1.2Covariance

Covariance is a measure of the joint variability of two random variables. It shows how changes in one random variable correspond to another changing variable. The covariance of a population is calculated based on equation (3).

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n} \qquad (3)$$

Where:

x_i, the ith value of x;

x, the mean of all x values;

y_i, the yth value;

$\bar{y}$, the mean of all y values; and

n, the total number of data.

2.4.1.3Variance

The variance shows the distance of each variable in the data set from the mean. It can be defined as a measure of the spread among the data. It is calculated based on equation (4) :

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} \qquad (4)$$

Where:

$\sigma,$ standard deviation;

σ^2, variance;

x_i, the ith value of x;

x, the mean of all x values; and

n, the number of data point.

Instead of calculating the covariance and variance separately and assembling the covariance matrix based on these values, the covariance matrix can be calculated through matrix operations. There are three steps to calculate the covariance matrix based on the original data matrix, as follows:

1. Center the data matrix X by calculating the mean of each column and subtracting the mean from the elements of each column;
2. Multiply the centered matrix by its transpose; and
3. Divide the results by n-1 where n is the number of rows;

The three aforementioned steps for calculating the covariance matrix, C, can be formulated as equation (5):

$$C = \frac{XX^T}{n-1} \tag{5}$$

Where:

X, centered data matrix;

X^T, transpose of centered s data matrix; and

n, number of rows of matrix X.

2.4.1.4Correlation matrix

The correlation matrix is equivalent to dividing each covariance matrix element by its standard deviation. It shows the relation among sets of given variables. One well-known form of the correlation matrix applies the Pearson correlation. The Pearson product-moment correlation can be calculated as equation ((6)). In this correlation, elements of the matrix ranged between -1, 1. Greater absolute values indicate a stronger relationship between the two variables. The sign represents their direct or inverse relationship. Accordingly, values close to zero show no correlation.

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \tag{6}$$

Where:

r_{xy}, Pearson correlation coefficient between variable x and y;

2.4.1.5Eigenvalue and Eigenvector

For any given matrix A, there exists a vector B and value λ, which satisfy the following equation:

$$AB = \lambda B \tag{7}$$

In this equation, λ is the eigenvalue, and the vector B, the eigenvector. In other words, for any n*n matrix A, the vectors B, and scalar values λ, which provide a solution to (7) are named eigenvectors and eigenvalues, respectively.

The eigenvectors, B, of the covariance matrix A, are the principal components, which show the direction in which most information is available. That is the direction in which most variance can be explained. The related eigenvalues represent the variance carried by each principal component. A key limitation of PCA is that PCs are not easily interpretable.

2.4.1.6Singular Value Decomposition

The Singular Value Decomposition is based on the fact that a matrix can be decomposed into two orthogonal matrices and a diagonal matrix. This has a direct mathematical relation to the PCA calculated using the covariance matrix. The SVD can indeed provide more information than the PCA calculated through the covariance matrix due to rounding. This extra information can be useful in applications where the structure of columns is as important as rows. In fact, the additional information provided through one of the decompositions of the data matrix, U. It can also detect and extract small signals from noisy data and handle sparse matrices. Moreover, achieving PCs through SVD is more efficient than calculating the eigenvalue and eigenvector of the covariance matrix (Jolliffe, 2002).

Singular Value Decomposition (SVD) states that any given matrix can be factorized as in equation (8)

$$X = USV^T \tag{8}$$

Where:

X, given m×n data matrix;

U, orthogonal m×m matrix;

S, diagonal m×n matrix; and

V, orthogonal n×n matrix;

Columns of U are called left singular vectors of X, and columns of V are right singular vectors of X. The matrix S is the scaling matrix, in which the diagonal entries are known as the singular value of X. For each singular vector, there is a corresponding unique singular value. These values are generally sorted in descending order within the scaling matrix.

According to Jolliffe (2002), it can be shown that when the data matrix is column centered, the right singular vectors are similar to the eigenvectors of the covariance matrix A, containing principal components. However, when the data matrix is row-centered, the left singular vectors are principal components (Wall et.al., 2003). The singular values can also be used to derive the eigenvalues of the covariance matrix. If the singular value S is divided by n-1, it will be equivalent to the square root of eigenvalues of the covariance matrix. The mentioned factors can provide information on PCs' loading and standard deviation.

It is worth mentioning that for SVD or PCA, normalizing the data matrix is essential. There is no guaranty that all the features have the same scale, and considering these various scales is not the responsibility of these approaches. Avoiding normalization before SVD/PCA might provide unreal results by unduly assigning greater importance to variables with a larger scale.

The number of non-zero singular values is equal to the rank of matrix X. Matrix rank is defined as the maximum number of the independent linear subset of rows, or the maximum number of independent linear subsets of columns (Roughgarden & Valiant, 2015). Accordingly, if the data matrix has fewer observations than variables, then the rank of the matrix would be based on the number of observed entries rather than variables. This, in turn, would lead to a reduction in the number of acceptable eigenvectors and PCs.

2.4.1.7 Variable Loading

Loading, as related to PCA, describes the contribution of each variable in a particular PC. More significant loadings indicate the greater correlation between a specific PC and a specific variable. Loading is calculated by multiplying the related eigenvector and eigenvalue of the covariance matrix and can easily be related to singular value decompositions through equation (9)

$$L = V\frac{s}{\sqrt{n-1}} \tag{9}$$

Where:

L, loading matrix

V, right singular vector; and

S, singular value;

2.4.1.8Score of samples

The score of samples indicates the location of the observed data in the new coordinates defined by PCs. These scores are the product of loadings and centered observations.

2.4.1.9General Assumptions and Limitations of PCA

Principal components are ordered according to decreasing variance. The first PC contains the most information (Jolliffe & Cadima, 2016).

There are three general assumptions in applying PCA (Shlens, 2014):

1. Since PCA works with Pearson correlation coefficient and covariance matrix, which are valid in the context of linear algebra, linear relationships between target variable and predictors is essential;

2. Large variances have important structure, meaning those with lower variance can be considered noise. Since not all the PCs are equally important, and the idea behind this method is dimensionality reduction with finding the least number of PCs, the variance carried by each PCs is defined as a measure of importance. The more variation along with each PCs, the more interesting and important PCs are; and

3. Principal components are independent. This assumption ensures that loading vectors of PCs are orthogonal, which facilitates finding the direction of PCs.

Since PCA is based on variance, which changes according to the magnitude of the variable, different units of measurement can change the covariance matrix. Accordingly, data standardization before conducting a PCA is an important point recommended by different authors to ensure all the data are on the same scale. Jolliffe and Cadima (2016), suggest that data should be centered, i.e., the mean should be subtracted from each attribute, particularly if SVD will be applied. Also, in the cases of non-linear relationship among data set that might contain all types of variable such as categorical, numerical, etc. standard PCA is not applicable, and an alternative approach or some modification of the standard PCA should be applied instead Jolliffe and Cadima (2016) and Jolliffe (2002). Non-linear principal component analysis is a modified PCA which is recommended for mixed data type that are not linearly dependent. This approach is explained in more detail in the following paragraphs.

2.4.2 Non-Linear PCA

Non-Linear PCA, also known as categorical PCA, is a dimensionality reduction method that, unlike PCA, can handle a non-linear relationship among variables. This method applies to datasets with different variable types such as nominal, ordinal, and numerical. Although the method follows the same objective as PCA, the detail of computing the outcomes are different. Results of the technique include eigenvalues, component loading, component score, and communality which describes Variance Accounted For (VAF) by each component, correlation among variables and components, the contribution of selected variables in total VAF, and the score of components associated with each case in dataset respectively (Linting & van der Kooij, 2012). In non-linear PCA, categories of variables are replaced with numerical values through a process called optimal scaling. In fact, the assigned numerical values indicate category quantifications. In non-linear PCA, the correlation among quantified variables is computed rather than that of the observed variables. Accordingly, the correlation matrix may change according to the type of assigned quantification (Mitsuhiro & Yadohisa, 2018).

2.4.2.1Optimal scaling

The optimal scaling method deals with variables in three different ways called analysis level. These levels are nominal, ordinal and numerical analysis levels and determine how category values can be transformed to category quantification (Linting & van der Kooij, 2012). The analysis levels can be promoted by the target of analysis and are regardless of variables properties. Optimal scaling follows the target of optimizing the correlation matrix of quantified variables.

The three mentioned analysis levels are explained briefly in the following paragraphs.

The Nominal analysis level applies to nominal categories such as pipe material or pipe ID in which their value and order do not convey specific meaning. In this analysis level, for variable j, if two separate observations i and j belong to the same category, i.e. $y_{ji} = y_{jh}$, their category quantification will be the same.

The ordinal analysis level applies to categorical variables in which values follow a specific order. Pipes condition, level of corrosion, etc., are some examples of ordinal data. In this analysis level, quantifications order needs to follow the same order of original categories, i.e., for $y_{ji} >_{yjh}, y^*_{ji} \geq y^*_{jh}$.

Numeric analysis level indicates groups of numerical values which are measurable. In fact, mathematical operations on these values are meaningful. Age of pipes, number of breaks, etc., are some examples of this group. In this analysis level, the vector of y_j is standardized and replaced with y^*_j with zero mean and unit variance.

In order to understand the optimal scaling process better, let us define a data matrix Y with two categorical variables, pipe material, and coating status. Each variable y_j contains K_j categories. Y can be written as below:

$$Y = (y_1, y_2) = \begin{pmatrix} Plastic & No \\ Plastic & No \\ Metal & Yes \\ AsbestosCement & No \\ Metal & Yes \end{pmatrix}$$

The first variable has three categories in this example matrix, and the second variable has two categories. Accordingly, K_j for each is three and two, respectively.

For each variable an indicator matrix, G_j, is defined. G_j is a $N \times k_j$ matrix of dummy variables. Elements of this matrix, i.e. g_{jik}, are zero when a certain object is not of the given category k, or one when it is of that category. This relation is defined as below (Yuichi et.al., 2016):

$$g_{jik} = \begin{cases} 1, & if\ object\ i\ belongs\ to\ \ category\ k \\ 0, & if\ object\ i\ does\ not\ belong\ to\ category\ k \end{cases}$$

Each categorical variable y_j Y is the inner product of a category vector and an indicator matrix in the data matrix.

For the given data matrix Y, the indicator matrix for each variable would be defined as below:

$$G_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, G_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Accordingly,

$$y_1 = G_1 \begin{pmatrix} Plastic \\ Metal \\ AsbestosCement \end{pmatrix}, \qquad y_1 = G_2 \begin{pmatrix} No \\ Yes \end{pmatrix}$$

The target of optimal scaling is to transform variable categories into numerical values called quantifications, i.e. q_j, thereby defining a new numerical variable.

The new transformed variable y_j^* is defined as (1010101010).

$$y_j^* = G_j \times q_j \tag{10}$$

Where:

G_j, binary indicator matrix for jth categorical variable; and

q_j, vector of category quantification.

To do so, homogeneity analysis with some restrictions is introduced as a loss function. Minimizing the loss function is a solution for finding q_j . Homogeneity analysis finds quantification of categories for each variable to maximize homogeneity (Yuichi et.al., 2016). The loss function indicating deviation from homogeneity is defined as (11) (Yuichi et.al., 2016):

$$\sigma_H(Z,W) = \sum_{J=1}^{P} tr(Z - G_j W_j)^T tr(Z - G_j W_j) = \sum_{j=1}^{p} \sigma_{Hj}(Z,W_j) \tag{11}$$

Where:

σ_H, Loss function indicating homogeneity deviation;

tr, the trace of a matrix defines as the summation of main diagonal elements;

r, number of components;

p, number of categorical variables;

T, transpose of a matrix

Z, N×r matrix of component scores; and

W_j, $k_j \times r$ matrix of category quantifications for variable j;

To find the solution, the equation must be minimized over Z and W considering restrictions (12) (Yuichi et.al., 2016):

$$Z^T 1_n = 0_r \text{ and } Z^T Z = nI_r \tag{12}$$

1_n, $N \times 1$ vector of ones;

0_r, $r \times 1$ vector of zeros; and

I_r, the identity matrix.

Using homogeneity in the concept of non-linear PCA add additional constraints on W_j as (13) (Yuichi et.al., 2016):

$$W_j = a_j \times q_j \tag{13}$$

Where:

a_j, is $1 \times r$ matrix of component loading.

To solve the minimization problem mentioned in eq (11), the loss function must be minimized for each categorical variable separately. To do so, an iterative process called Alternating Least Squares Algorithm is performed. Yuichi et.al., (2016) defined this process as well as the initial estimation for starting the process as follows.

Before starting the iterative process, initial values of Z and W_j need to be determined for each variable separately. The first estimation of Z, i.e., Z^0, is selected randomly considering restrictions mentioned in eq.12. According to this value, the first estimation of W_j^0 is calculated based on equation (14) (Yuichi et.al., 2016):

$$W_j^{(0)} = (G_j^T G_j)^{-1} G_j^T Z^{(0)} \tag{14}$$

Accordingly, q_j^0 determines as k_j successive integer variable based on analysis level. Then, the first estimation of the loading vector of a_j is calculated as (15) (Yuichi et.al., 2016):

$$a_j^{(0)} = Z^{(0)T} G_j q_j^0 \tag{15}$$

After computing the initial values, an iterative process consisting of four steps is performed as below:

1. Estimating the category quantification (W_j) for each variable according to equation (16) (Yuichi et.al., 2016):

$$W_j^{(t+1)} = (G_j^T G_j)^{-1} G_j^T Z^{(t)} \tag{16}$$

Where:

t, is the number of iterations.

2. Updating loading vector a_j as equation (17) (Yuichi et.al., 2016):

$$a_j^{(t+1)} = W_j^{(t+1)}(G_j^T G_j)q_j^t / a_j^{(t)T}(G_j^T G_j)q_j^t \tag{17}$$

3. Computing q_j for nominal variables as equation (18) (Yuichi et.al., 2016):(17)

$$q_j^{(t+1)} = W_j^{(t+1)} a_j^{(t+1)T} / a_j^{(t+1)} a_j^{(t+1)T} \tag{18}$$

4. Updating object score Z according to equation (19) (Yuichi et.al., 2016):

$$Z^{(t+1)} = \frac{1}{p} \sum_{1}^{P} G_j W_j^{t+1} \tag{19}$$

The values then are checked in the concept of minimizing the loss function mentioned in (11).

This iterative process continues until the results of the loss function converge.

2.4.3 Factor analysis of mixed data (FAMD)

Factor Analysis of Mixed Data is a dimensionality reduction approach which is employed to explore data with both continuous and categorical variables. This method is the combination of PCA and Multiple Correspondence Analysis (MCA). Accordingly, numerical variables are scaled to unit variance and handled in the same way as PCA. Also, Categorical variables are transferred into a Complete Disjunctive Matrix (CDM) and scaled according to the MCA-specific scaling. PCA is explained in-depth in 2.4.1 and MCA will be explained in 2.4.3.1.

To better understand the FAMD analysis let's assume we have I individual with $K_1 = \{1, K_1\}$ numerical variables and Q categorical variables $\{q = 1, Q\}$ in which each variable q has K_q categories, and the total number of categories is $K_2 = \sum_q K_q$.

In FAMD, the original data with I individual and $K_1 + Q$ variables are defined in the form of a data matrix, X, with I individual and $K = K_1 + K_2$ columns. The elements of matrix X are standardized based on PCA and MCA rules for numerical and categorical variables respectively. Accordingly, in FAMD analysis data matrix X tries to project these data in direction, v, which maximizes the variance. Therefore, the target of FAMD can be formulated as maximizing the following (Pagès, 2014):

$$\sum_{k \in K_1} r^2\,(k, v) + \sum_{q \in Q} \eta^2(q, v) \qquad (20)$$

Where:

r, the correlation coefficient between variable k and vector v; and

η, correlation ratio between variable q and vector v.

The first element of equation (20) is PCA optimization problem and the second is related to MCA.

2.4.3.1Multiple Correspondence Analysis

Multiple Correspondence Analysis (MCA) is a principal component method that analysis the patterns and relationships between categorical variables. This method is a special case of PCA (Pagès, 2014), in which a set of Q qualitative variables are replaced with some quantitative variables through assigning a coefficient to each category. In MCA, a dataset with I individual and Q set of the qualitative variable with K_q categories can be represented in form of a data table called

Complete Disjunctive Matrix (CDM). As mentioned earlier the total number of categories is $K_2 = \sum_q K_q$

CDM is a data matrix with I individuals in the rows and categories of variables in the columns as such elements in the intersection of row i and column K_q of variable q, i.e. y_{ik}, are zero when a certain object is not of the given category, or one when it is of that category. As in PCA, to implement MCA, the elements of CDM should be standardized Jérôme Pagès (2014). The standardization of CDM elements in MCA analysis can be formulated as equation (21) (Pagès, 2014):

$$x_{ik} = {}^{y_{ik}}/_{p_k} - 1 \qquad (21)$$

Where:

y_{ik}, elements of CDM in the intersection of row i and column K_q of variable q; and

p_k, is the proportion of individuals in category K_q.

According to data matrix X, each column K_q takes one value per individual. Meaning that, in column K_q, the x_{ik} have I dimension which each dimension corresponds to an individual. The variance of a category K_q is defined as the square distance of that category from the origin and can be formulated as (22) (Pagès, 2014):

$$Var(k) = \frac{1}{p_k} - 1 \qquad (22)$$

The mentioned formula indicates that the rare categories have a higher variance. The rare categories are located far from the origin. Since in principal component methods, inertia is more important than the distance, the total inertia of a variable q can be formulated as (23) (Pagès, 2014):

$$Inertia(q) = \frac{K_q - 1}{Q} \tag{23}$$

The mentioned formula reveals that inertia of variable q is proportional to the number of categories it has minus 1. Accordingly, the inertia of a variable with, let's say, 21 categories is 20 times more than a variable with 2 categories. Therefore, the variables with fewer categories represent fewer dimensions and vice versa.

The target of MCA, similar to PCA, is to project the point clouds of into smaller dimensional space while maximizing inertia. To do so, the clouds should be projected into a set of orthogonal axes with maximal inertia (variance). When projecting all categories of variable q, denoting by K_q, on a unit vector ν, the inertia of ν is one. Hence, the between-class inertia, i.e., the inertia explained by each category of variable q, can be defined as a percentage of inertia of ν explained by qualitative variable q. This percentage is equal to the squared correlation ratio between qualitative variable q and quantitative variable ν and can be formulated as equation (24) (Pagès, 2014):

$$Between\ class\ inertia = \frac{1}{J}\varphi^2 \tag{24}$$

Where:

φ, is correlation ratio between variable q and component ν

The percentage of inertia indicates the quality of representing variables by a given axes (Pagès, 2014). The percentage of inertia in MCA is smaller compared with PCA, and the maximum percentage of variance for categories of variable j, cannot exceed $\frac{1}{K_j - 1}$ (Pagès, 2014). A squared correlation ratio is employed to measure how the components are related to the variables.

2.4.3.2MCA and PCA relationship

MCA looks for the principal components that are related to the variables as much as possible. Defining eigen value λ_s as an intensity measurement of relationship between variables the following equation (25) can be proved (Pagès, 2014):

$$XM\acute{X}Dv_s = \lambda_s v_s \tag{25}$$

Where:

X, transformed CDT matrix

M, Diagonal matrix of with $\frac{P_k}{Q}$ in the diagonal elements;

D, Diagonal matrix with P_i in diagonal elements;

λ_s, eigen value of $XM\acute{X}D$ corresponds to the projected inertia; and

v_s, Eigenvector of $XM\acute{X}D$ corresponds to the direction in which variance is maximized.

Comparing MCA with PCA indicates that the process of MCA is similar to PCA with additional categories weight, M.

2.4.4 Automatic variable selection

Automatic variable selection can be considered a dimensionality reduction approach. However, unlike the other methods previously mentioned, it requires choosing a model before variable selection. The following paragraphs explain some popular methods of this approach.

2.4.4.1Stepwise regression

Stepwise regression is one of the well-known methods of automatic variable selection. This method works with regression models such as logistic and linear regression. After choosing a regression model, a precision criterion is used for evaluating the significance of independent variables. Accordingly, predictive variables are added or removed from the analysis according to their level of statistical significance (Frost, 2020). At the end of the stepwise process a single regression model is produced. Potential precision criteria include R^2, Residual sum of squares (RSS) and Akaike Information Criterion (AIC). Due to the large space of possible models that stepwise regression searches within, the possibility of overfitting increases. Accordingly, using criteria such as AIC to detect overfitting is recommended. The measurement criteria in section 2.4.4.4 provides a brief explanation of the mentioned criteria.

The stepwise regression algorithm is a combination of Forward selection and Backward elimination, meaning that in each step, a variable is added, and all the selected variables in previous steps are checked whether their significance level is changed and reduced below the tolerated level. In case of finding nonsignificant variable, it will remove from the analysis (Stepwise regression).

2.4.4.1.1 Forward selection

Forward selection is an iterative process that starts by considering no variables and evaluating the results of adding additional variables, one at a time, according to a specific model fit criterion. The process continues until the remaining variables do not significantly improve results. Selected variables are those with remarkable impacts on the fit of the model.

2.4.4.1.2 Backward elimination

Unlike forward selection, in backward elimination, all available variables are considered first. In the following steps, variables are iteratively removed, as long as they do not highly affect the model's fit.

2.4.4.2Best subset regression

Best subset regression considers all possible regression models based on independent variables. Thus, with p independent variables, 2^pmodels are created. Each set of with the same number of variables are compared according to fit criteria such as R-squared or C_p. Therefore, the results represent the best models of different variable subsets ranging from 1 to p. The best of the best is then chosen as the best subset regression model (Karimian, 2016)

Although best subset regression analyses more combinations of regression models and can lead to better results, stepwise regression is more straightforward (Frost, 2020).

2.4.4.3Recursive Feature Elimination (RFE)

Recursive feature elimination (RFE) is a feature elimination method in which the most important factors are determined through a backward elimination process. RFE initially introduced by Guyon et.al. (2002). This method represents the ranking of features and subset of data with the corresponding accuracy. The selected candidates are the predefined number of predictors or are a subset of the variables corresponding to the highest resulting accuracy (Chen et.al., 2018). RFE method is basically employed along with classification algorithms such as random forest and support vector machine (SVM) (Chen et.al., 2018) as well as regression algorithms including but not limited to linear regression, logistic regression, etc. as an estimator (Kuhn & Johnson, 2019). When the number of required variables is defined in advance for RFE, the top N variables will be selected according to the criteria specified by the estimator through a backward elimination process. The chosen subset corresponds to the highest accuracy. Selection of data solely based on the highest pre-defined number of variables might lead to selecting a large subset of data. Also, the pre-selection number of predictors without prior knowledge might lead to biased results (Chen et.al., 2018). To overcome this issue, RFE methods are usually combined with some other "decision variant" approaches to determine the optimal number of variables from the selected subset of variables based on the accuracy (Chen et.al., 2018). Imagine a dataset with a P number of features. Recursive Feature Elimination with Cross Validation is a well-known RFE approach with a decision variant in which the best subsets of features are selected through specified estimator by removing 0 to P features through the RFE process. The optimal best subset is then chosen according to the model's cross-validation score. The cross-validation process is briefly explained in section 2.4.4.3.2.1. In general, the effectiveness of the RFE approach depends on two factors: the estimator combined with this method and the performance evaluation criteria. Hence, selecting a proper estimator according to the type of attributes and the number of data is one of the key steps of the RFE approach. As mentioned earlier, RFE selects the most important features through an estimator. An estimator is a supervised algorithm with a fit method that is employed to identify important features. Depending on the attribute target, various algorithms can be used for this purpose. A study by Granito et.al., (2006) on identifying the most important features in the producing spectra through Proton transfer compared random forest and logistic regression as two estimators for the RFE process. The analysis results indicate that the random forest recursive feature elimination (RF-RFE) outperforms logistic regression recursive feature elimination (LR-RFE). The higher performance of RF-RFE compared with other estimators such as linear Support Vector Classification (linear-SVC), Extra-Tree Classifier, AdaBoost, and naïve bias is also confirmed in the results of other studies such as Fang et.al., (2020), and Wang and Chen (2019). Besides random forest, Extreme Gradient Boosting (XGBOOST) approach has been performed well as an estimator of RFECV and as a prediction model.

Results of a study by Chang et.al., (2019), employed four different supervised algorithms, i.e., Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBOOST), as an estimator of RFECV approach to identify the most relevant features in predicting the outcome of hypertension. In the same study, the same supervised algorithms were employed to predict hypertension based on the selected features. The study results indicate that RF, DT, and XGBOOST outperformed SVM in selecting the important features and in the next step, among all the four prediction models, XGBOOST has performed more accurately than the others. Based on the mentioned literature, for this study, random forest and XGBOOST are selected as an estimator of the RFECV approach for finding the most important features. Also, to ensure the effectiveness of these approaches, the performance of the approaches on predicting the two target attributes with the selected features will be evaluated. More detail of random forest and XGBOOST is provided in 2.4.4.3.1 and 2.4.4.3.2, respectively.

2.4.4.3.1 Random Forest

Random forest is a supervised learning approach used for classification and regression analysis and consists of multiple classification trees. This method creates uncorrelated forests from decision trees, classifying the target variable more accurately than an individual tree. Each decision tree consists of multiple nodes. Features in the nodes decide how a dataset should be divided into two sub-classes. In this method, internal features are selected to decrease the impurity of the selected classes based on specific criteria. For classification, impurity is generally based on Gini impurity or entropy, whereas regression is based on variance reduction. Each of the mentioned criteria is briefly explained in the following paragraphs. Attributes with the most significant impurity decrease are selected as internal nodes and are assigned the highest weights. In this approach, the final prediction is based on the average of all trees or majority votes for classification and regression. Since this method is built from multiple trees, with random predictors, it can assign weights to each feature and determine the important features while making the prediction (Cooper et.al., 2012). Multicollinearity does not affect Random Forest results. The graphical explanation of this approach is provided in Figure 2-1.

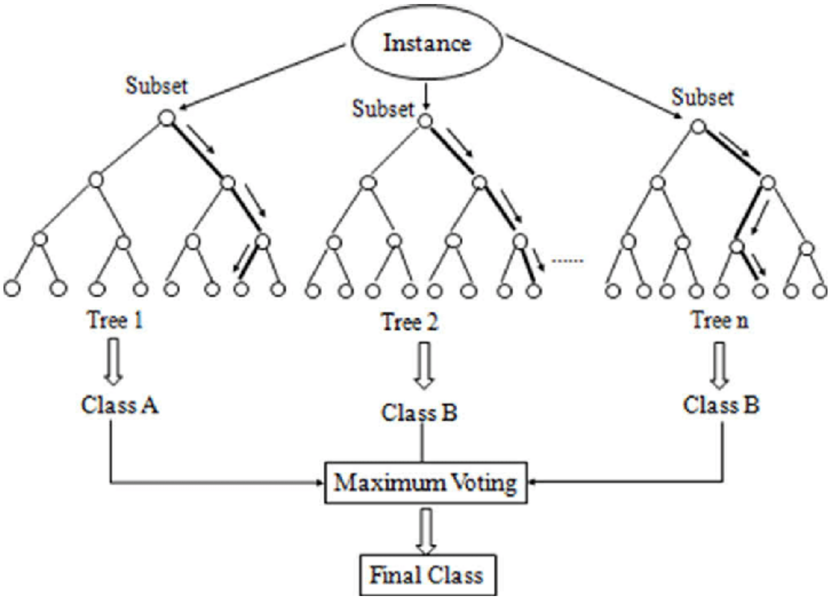


Figure 2-1 Random Forest structure (Nain et.al., 2018)

2.4.4.3.1.1 Gini impurity

Gini impurity measures how likely the incorrect classification of a new instance of a random variable is when randomly classified (Ambielli, 2017). This criterion for a target variable with C classes can be formulated as equation (26).

$$Gini = 1 - \sum (p_i)^2 \tag{26}$$

Where:

C, Number of classes; and

p(i), probability of picking a datapoint with class i.

2.4.4.3.1.2 Entropy impurity

Entropy impurity measures how much variance is in data. This measurement can be formulated as equation (27):

$$Entropy = -\sum p_i \; . \;\; log_2 p_i \tag{27}$$

2.4.4.3.2 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a supervised machine learning algorithm suitable for regression and classification problems, and similar to RF can identify the most relevant attributes to the target. In XGBoost, the main focus is on combining simpler and weaker models' estimates to predict a classification or regression target accurately. More specifically, several trees are combined and trained the model (Chen & Guestrin, 2016). The weak learners in gradient boosting for regression are regression trees. Each regression tree maps an input data point to one of its leaves containing a continuous score. XGBoost combines a convex loss function with a penalty term for model complexity to minimize an objective function (in other words, to minimize the regression tree functions). The convex loss function is the difference between the predicted and target outputs. The training process is repeated iteratively, with new trees being added that forecast the residuals or errors of previous trees, which are then integrated with previous trees to provide the final prediction. In other words, XGBoost is an iterative process in which the residuals are calculated during each iteration, and proceeding predictors are adjusted to optimize a particular loss function. According to Snider and McBean (2020), XGBoost has a shorter training process and is robust against noise and outliers. In addition, as mentioned in the literature, this approach has a high prediction performance compared with other machine learning algorithms.

2.4.4.3.2.1 Cross-validation

Cross-validation is a sampling approach employed to evaluate machine learning models. The main parameters of this technique are the number of groups required for splitting data, i.e., k, and evaluation metric. The data is divided into k subsets in this approach, and the process iterates k times. One subset is considered a test set in each trial and the other k-1 subsets as the training set. The model is learned from the k-1 subsets and then is tested out on the test set. In each trial, the analysis score on the test set is evaluated. The overall performance of the model is the average of the k scores. For this study, k-fold cross-validation is employed along with randomized search and RFE methods for tunning the RFE estimator and optimally finding the most important features, respectively. Also, cross-validation is used alone to ensure the tuned estimator of the RFECV method is not overfitted.

2.4.4.4The goodness of fit measurement criteria

2.4.4.4.1 R-Squared

R^2 is a statistical measurement for evaluating the fit of regression models. This evaluation criterion indicates the amounts of variance related to the dependent target variable that independent variables can explain. Higher values of R^2 indicate a better fit of the model. In order to calculate R^2, first, the line of best fit should be found based on the data points of dependent and independent variables. Next, predicted values should be subtracted from actual values and squared. The result is a list of squared errors summed to indicate unexplained variance. The average of actual values should be subtracted from each actual value, squared, and summed to calculate total variance. R^2 is then calculated by dividing unexplained variance by total variance and subtracting results from one. The expression can be formulated as (28):

$$R^2 = 1 - \frac{unexplained\ variance}{Total\ variance} \tag{28}$$

R^2 can also lead to overfitting as a high value of it can be erroneously interpreted as better. In fact, R^2 does not provide information on whether data and predictions are biased or model goodness (Hayes, 2020). Additionally, if more predictors are added to the analysis, R^2 naturally becomes higher, skewing results. Accordingly, this criterion is more useful when comparing models of the same size.

2.4.4.4.2 Residual Sum of Squares (RSS)

RSS is a statistical technique for evaluating the precision of regression models. In a dataset, the amount of variance that regression models do not explain can be measured by RSS (Barone, 2019). This criterion is the sum of the squared differences between actual and predicted values. Accordingly, smaller values of RSS are preferred. However, as in R^2, considering more variables in the analysis can improve the results of RSS. Hence, this method criterion is prone to overfitting and is also recommended for comparing models with the same number of predictors. It is important to note that RSS values depend on the unit of measure because it is not normalized.

2.4.4.4.3 Mallows C_p

Mallows C_p is a measurement criterion for evaluating the fit of multiple regression models. This measurement compares the precision of full models with smaller predictors and determines the amount of unexplained error by model. The smaller values of C_p is desirable (Stephanie, 2020). It is noticeable that this criterion can address the issue of overfitting and can be formulated as equation (29):

$$C_p = \frac{RSS}{s^2} + 2p - n \tag{29}$$

Where:

RSS, Residual sum square explained in section 2.4.4.4.2 for the model with p-1 variable;

s^2, Residual mean square for the model with all available variables;

p, number of the variable used for model plus one; and

n, number of observations.

2.4.4.4.4 Akaike information criterion

AIC is a measure of evaluating the fit of a statistical model and indicates the amount of information lost by the model. Accordingly, lower values of AIC indicate that the model is well fit to the data and not overfit (Date, 2019). This criterion can also detect overfitting. This measurement is formulated as (30):

$$AIC = 2P - 2\ln\hat{L} \tag{30}$$

Where:

P, number of variables considered in the model; and

$\hat{L}$, the maximum value of the likelihood function of the model.

2.4.4.4.5 Likelihood function

The likelihood function indicates the possibility of observing different observations according to the dataset. Finding an optimal way to fit a distribution in a dataset is the target of the maximum likelihood function. In other words, for the distribution of the observation, we are looking for the optimal value of mean or standard deviation.

2.4.4.4.6 F1-score

F1 score is an evaluation metric for categorical targets. This measurement is the harmonic mean of precision and recall and can be calculated based on equation (31):

$$F1 = \frac{2(Precision * Recal)}{Precision + Recall} \tag{31}$$

Precision (Pr) indicates the percentage of the positive identifications that were positive originally, and Recall (Re) represents the proportion of the actual positives identified correctly. The two mentioned criteria can be formulated as (32) and (33), respectively.

$$Pr = \frac{TP}{TP + FP} \tag{32}$$

Where:

TP, indicates the number of values that are actually positive and are predicted as positive; and

FP, indicates the number of values that are actually negative and incorrectly are predicted as positive.

$$Re = \frac{TP}{TP + FN} \tag{33}$$

Where:

FN, indicates the number of values that are actually positive and incorrectly are predicted as negative

Comparing the mentioned criteria indicates, AIC criterion performs well for stepwise regression, and Mallows C_p is recommended for best subset regression.

Since in this study the data set is a combination of numerical and categorical variables and also depending on the mentioned literature, this study has been focused on four main models: Factor analysis of mixed data (FAMD), categorical PCA (CATPCA), random forest recursive feature elimination with cross-validation (RF-RFECV), and extreme gradient boosting forest recursive feature elimination with cross-validation (XGBOOST-RFECV).

3 Methodology

The analysis of important factors driving watermain failure was divided into five key steps, data understanding, data cleaning, data preparation, correlation analysis, and applying different dimensionality reduction approaches. Each is explained in more detail in the following paragraphs.

3.1 Data understanding

Understanding the data is a key preprocessing step that helps identify the required data cleaning process and select appropriate analysis methods.

This process began with the creation of metadata tables to characterize the available raw data and ensure that all information from primary datasets would be accurately interpreted in the subsequent steps. These metadata tables reflect the following information:

Title: Actual name of the attribute provided in the original data set.

Description: A brief explanation of the attribute to better understand the information provided by the attribute.

Type: represents information on the kind of a specific attribute. Overall, the variables in this study are categorized into groups of numerical, categorical, and polynomial.

Name: A defined unique name for each attribute that is the same along with all utilities. Since different utilities recorded the same information into a different name, defining this unique name for better identifying similar attributes is essential.

Unit: Related unit of the numerical variables. For instance, for diameter, the unit is mm.

Category: Revels different categories of categorical variables.

Range: Domain of the numerical variables.

With the provided metadata tables, the inconsistencies and problems in the data, unique IDs for further matching break and inventory data, and the most appropriate dimensionality reduction methods were identified. A sample of the mentioned metadata table is provided in Table 3-1.

Table 3-1 Sample of metadata table for the city of Calgary

Title	**Description**	**Type**	**Name**	**Unit**	**Range**	**Category**
BREAK_APPA	Cause of break	Categorical	BreakCause	-	-	Age Corrosion Frozen soil
REL_WATMAI	Related pipe ID	Categorical	PipeID	-	-	1-991853
INCIDENT_D	Break date	Date	BreakDate	year	1986-2018	-
PIPE_SIZE	Nominal pipe diameter	Numerical	Diameter	mm	25-1200	-

3.2 Data cleaning

The data cleaning process includes two key steps: 1. Defining unique values among the 13 utilities and removing outliers, and 2. filling the gaps in data. The majority of the attributes required cleaning. Each of the mentioned steps is briefly explained in the following paragraphs:

3.2.1 Defining unique values and removing outliers

The collected data by each utility was recorded in various formats. Depending on the type of attribute, a different cleaning process was required. For categorical variables in general, unique names were defined for different variable categories. For instance, the pipe material for one utility was coded or abbreviated, while in the other utilities, it was the full name of that material. A full name of material was defined for each category of material. The cleaning process of numerical variables includes keeping values of a variable in the same units and removing outliers such as zero diameters or negative years from the data. Accordingly, the unique values were defined within the utilities to achieve an array of consistent datasets for further efficient data analysis. The cleaned values and the original dataset were imported to PowerBI to replace the cleaned values with the original ones. The defined structure of the different attributes is provided in Appendix E

3.2.2 Filling the gaps

Consistency in a dataset and lack of missing values are key terms in developing reliable deterioration and prediction models. Missing values commonly observed in various research usually arise from a lack of proper technological and financial resources. This incomplete information limits the applicability and interpretation of data, leading to information loss, false, and biased results. Hence, handling them in the initial and preprocessing steps of the analysis is essential. This study employed four main approaches to fill the gaps and missing values in the data set. The methods were: 1. Assuming a value, 2. Mirroring attributes, 3. Homogeneity analysis, and 4. Adjacent assets. The applicability of the proposed methods depends on the type of attributes, logical patterns in data, and available information. While the homogenous groups' method is applicable in various cases, the mirroring attribute and assuming a value depend on the type of attribute and available information. Also, and adjacent assets required the availability of GIS information. Each of the mentioned approach are briefly explained in the following paragraphs.

3.2.2.1Assumed value

Assuming a value for filling the gaps in a dataset assigns a value to the missing values based on some logical assumption and expert rules. This approach is recommended for binary attributes where a logical pattern is observed in the data. Assume value might not be advanced but can be the starting point for quickly filling the gaps. For instance, since the anode status of pipes with anode protection is “Yes”, it can be assumed that the missing values are related to pipes without anode protection.

3.2.2.2Mirroring attribute

In a dataset, some attributes might reflect the same information but in a different level of detail. Mirroring attribute is a method in which information from similar attributes is used for replacing the missing values.

3.2.2.3Homogeneous groups (Statistical measure)

Homogeneity analysis groups data based on similar characteristics. Assets within the same group are expected to have similar specifications. Missing values can be replaced based on the statistical characteristics of the group. For instance, Over the years, improvements in pipe manufacturing processes of different materials have led to changing trends in material use; A study by Kirmeyer (1994) estimated in 1992 that DI and CI covered two-third of watermains in the USA, AC pipes around 15% and the remaining either plastic or concrete. However, during recent years this pattern is changed. Accordingly, one can assume a value for material based on the material mode in the same installation year.

3.2.2.4Adjacent assets

Adjacent assets mainly focus on replacing missing values based on information of near assets. Some nearby assets are expected to have similar characteristics such as install year, diameter, etc. The adjacent assets can be identified through tools such as GIS.

The Table 3-2 indicates the techniques employed for filling the missing values in each attribute. It is noticeable that in the cases these methods were not applicable, the missing values were removed from the analysis.

Table 3-2 Methods for filling the gaps in each attribute

Missing value	**Method**	**Explanation**
Lining Status	Assume a value	When there is lining on a pipe, usually it is shown as Yes. It can be concluded that the missing values are related to the pipes without lining protection. This is employed when no information on mirroring attribute was available to fill the gaps.
	Mirroring Attribute	Missing lining status values replaced based on values of lining material. In cases that the lining material contains "UnLined" value, the lining status is replaced with No and Yes otherwise.

Missing value	**Method**	**Explanation**
Anode Status	Assume a value	When there is anode protection, usually it is shown as Yes. It can be concluded that the missing values are related to the pipes without Anode protection. This is employed when no information on mirroring attribute was available to fill the gaps.
	Homogenous groups	Pipes with the same material are expected to have the same Anode Status. Accordingly, for a missing anode status, material checked, and based on the mode of Anode Status for that material, the missing value replaced.
Material	Homogenous groups	The pipes categorized based on their install year to check the installation year of that pipe for missing material. Accordingly, the mode of material for that year picked as a predicted value for that pipe.
	Adjacent assets	Information from near pipes material used for replacing missing material for cities with available GIS.
Diameter	Homogenous groups	Missing diameter replaced by the median of diameter.
	Adjacent assets	Information from near pipes material used for replacing missing diameter for cities with available GIS.
Lining Material	Mirroring attribute	Information on lining status was employed for this purpose. Status of No indicates "Unlined" pipes.
Coating material	Mirroring attribute	Information on coating status was employed for this purpose. Uncoated pipes are related.
Install year	Homogenous groups	Mode the installation years for that specific material employed to replace the missing install year.
Depth	Homogenous groups	Median of depth values used for this purpose.

Missing value	**Method**	**Explanation**
Protection status	Assume a value	When there is protection on a pipe, usually it is shown as Yes. It can be concluded that the missing values are related to the pipes without protection.
Coating status	Assume a value	When there is coating on a pipe, usually it is shown as Yes. It can be concluded that the missing values are related to the pipes without protection.
	Mirroring attribute	Missing coating status values replaced based on values of coating material. In cases that the coating material contains an "Uncoated" value, the lining status is replaced with No and Yes otherwise. This is employed when no information on mirroring attribute was available to fill the gaps.

3.3 Data preparation

Four workshops were organized with National Water and Wastewater Benchmarking Initiative (NWWBI) to identify the current watermain challenges and the available repair and replacement strategies in utilities. During these meetings, two prediction targets were found to be preferable by the utilities: estimating the rate of failure and predicting the probability of failure. These targets can help utilities in risk management associated with asset management plans and estimating the periodical repair and maintenance requirements. Hence, the main focus of this study was defined to identify the factors affecting these targets. Accordingly, two target attributes have been defined Break status (categorical target) and current rate of failure (numerical target). While the first requires records of both broken and non-broken pipes, the second requires records of broken pipes only. The data preparation steps are explained in more detail in the following paragraphs:

3.3.1 Break status

The break status, determining whether a pipe will break or not, require information on both broken and non-broken pipes. Hence, the available inventory and break dataset were merged for each utility separately. The unique ID for each utility identified through metadata tables was employed to combine the two files in PowerBI. In the cases without a pipe ID in one of the files or low matching percentages, datasets were matched in GIS. The information related to the matching percentage of the two datasets is provided in Table 3-3. Comparison of the unique IDs in break and inventory indicates if a pipe has been failed or not. The broken pipes have similar IDs in the break and inventory file, while the ID of the non-broken pipe in the broken file is blank. For this target attribute age of the pipe is the difference between failure year and install year for the broken pipes, and for the non-broken pipes, the age is the difference between the most recent available failure year and install year. Finally, the break status target attribute and the list of attributes mentioned in Table 4-2 were analyzed to identify the most important factors leading to pipe failure.

Table 3-3 Matching percentage of break and inventory pipe based on Pipe ID

Utility	**% matching**	**Comment**
Barrie	98	-
Calgary	94	-
Durham	99	-
Halifax	99	-
Kitchener	86	-
Markham	88	-
Region of Waterloo	94	-
Saskatoon	100	Match in QGIS
St.John’s	100	Match in QGIS
Vancouver	100	Match in QGIS
Victoria	95	Match in QGIS
Waterloo	87	-
Winnipeg	100	Match in QGIS

3.3.2 Current rate of failure

For the current rate of failure target, the current number of failures indicates the number of failures that occurred in the latest year per length. This attribute is calculated by counting recurrent pipe IDs in the latest break year divided by the length of the pipes. Also, to identify the impacts of previous failures on the current rate of failure, the previous rate of failure was calculated as the proportion of all previous failures that occurred per length at each age. The age of the pipes for this target attribute is the difference between the current failure year and pipe installation year. Finally, the current rate of failure target attribute and the list of attributes mentioned in Table 4-1 were analyzed to identify the most important factors leading to pipe failure.

3.3.3 Converting categorical variables to numerical

Converting categorical variables to numerical ones is a crucial preprocessing step for some analyses, such as Recursive Feature Elimination (RFE) and correlation analysis. For the current study, the categorical variables converted to numerical ones through optimal scaling explained in the literature. Optimal scaling is known as opscale function available in optiscale library of Gifi package in R. General setups of opscale includes specifying the data frame of categorical variables (x.qual), the length of the data frame(x.quant), the measurement level of the variables (level), and measurement process (process). Measurement level can be nominal (1) or ordinal (2). For the ordinal values the measurement process can be either discrete (1) or continuous (2), and for the nominal values, there is no need to specify this option. For the current study, x.qual is the data frame of a categorical variable, and since the categorical data were all nominal, level 1 was selected as the analysis level.

3.4 Correlation analysis

In order to realize the relationship between the attributes and the targets, correlation analysis has been performed on data. Since correlation analysis reveals the relationship between numerical attributes, the converted categories were used in the analysis for categorical variables. A separate analysis has been performed on each target with common attributes and all other attributes recorded by utilities for each target.

Common attributes include diameter, age, length, and material available for all utilities. Initially, the common attributes of each target in all utilities were appended in Power BI. Then, the correlation analysis was performed on the appended dataset using python.

A separate correlation analysis has been performed for all other attributes in each utility. Then the list of correlation coefficients between each attribute and the target was recorded in a CSV file. Each column in the CSV file indicates the correlation coefficients between that attribute and the target in each utility. The number of correlation coefficient values for each attribute depends on the number of utilities that record that variable. The CSV file was then imported to python, and the range of correlation coefficients for each attribute was presented in a boxplot.

3.5 Dimensionality reduction approaches

As mentioned earlier, there are two types of dimensionality reduction approaches feature extraction and feature elimination. The present study applied two feature extraction methods and one feature elimination approach to identify the most important factors driving watermain failure. Based on the provided literature, type of data, as well as trial and error, FAMD and CATPCA were employed as feature extraction approaches, and Random Forest-RFECV and XGBoost – RFECV as feature elimination approaches for each of the target attributes. It is noticeable that the best subset regression and stepwise regression were tried on the data for one city for the regression problem. However, the resulting negative R^2 indicated these models are not a suitable feature selection for the data of this study.

The following paragraphs briefly explain the steps that have been taken for building each mentioned method:

3.5.1 Factor Analysis of Mixed Data (FAMD)

The FAMD was conducted in R programming language. This function is available in the FactoMineR library and requires setting the number of PCs as well as specifying the index of the target attribute. For this study, the number of PCs was set to equal the number of available predictors for each utility. In this analysis, the overall contribution of each variable in the PCs represents the significance of the variables. The contribution of variables in each PC and variance explained by each PC are the two key outputs of FAMD available in the Factoextra library. The overall contribution of each variable is the inner product of variables contribution in each PC and variance divided by total explained variance. The larger percentage of contribution indicates the more important factors.

3.5.2 Categorical Principal Component Analysis (CATPCA)

The CATPCA analysis was conducted in R. This function is known as princals and is available in the Gifi library. Princals mainly requires specifying the number of components (ndim), type of data (ordinal), location of knots for spline transformation, and degree of spline transformation. Initially, the number of required PCs was set to equal the total number of available predictors for each utility. Later, a proper number of components were selected depending on the amount of the explained variance. The selected PCs accounted for around 78-85% of the variance. Also, since the dataset includes numerical and categorical predictors, a degree vector was defined to properly transform categorical and numerical variables. Linear transformation of the numerical variables requires specifying the number of interior knots and degree of transformation 0 and 1, respectively (Linting et.al.,2007), and for the nominal transformation of the categorical variables, the degree was set to -1. Accordingly, categorical variables were handled at the nominal level, and numerical variables were linearly transformed. The key outputs of princals are the loadings of variables as well as the eigenvalues. Accordingly, the contribution of each variable in a specific principal component is the square of the variable loading associated with that PC divided by the sum square of all loadings of that component. The overall contribution of each variable is the inner product of variables contribution in each PC and the associated eigenvalue divided by the total eigenvalue.

3.5.3 Recursive feature elimination with cross-validation (RFECV)

The RFECV approach was conducted in a python Jupiter notebook. This function is available in the Scikit-Learn library. Since there was no prior knowledge about a proper number of desired features, RFE alone could not provide reliable results. Hence, the Recursive feature elimination with cross-validation (RFECV) approach was conducted instead. Initially, datasets were divided into training and test set using train_test_split available in sklearn.model_selection. The split ratio was selected between 0.09 – 0.3 depending on the number of data points as well as the fit of the model. Before the RFECV process started, a correlation analysis was performed on the training datasets for each city, and highly correlated predictors, i.e., correlation coefficient>0.8, were excluded from the analysis. The RFECV approach requires specifying several parameters in which the most important ones are briefly explained in Appendix A section A 1.

As mentioned earlier in 2.4.4.3, in this analysis, the most important features are selected by a supervised learning algorithm called "estimator". For this study, two estimators, i.e., random forest and XGBOOST, were identified to provide the best fit in the data. The best hyperparameters of the estimators were first tuned using RandomizedSearchCV available in Scikit-learn. Hyperparameters are parameters of a model set in advance to control the learning process of a machine learning algorithm. Since these parameters have a key role in fitting the model, optimizing them in advance provides more accurate and reliable results. These hyperparameters are different for each model. The most important hyperparameters for each estimator are described in Appendix A sections A 1 and A 2.

RandomizedSearchCV is a function that randomly picks values from the defined range of a model's hyperparameters and creates a combination of hyperparameters. The best combination is then selected through the cross-validation process explained in section 2.4.4.3.2.1. The possible number of hyperparameter combinations is determined by n_iter. A careful selection of n_iter is key to avoid missing the best parameters. 3-fold cross-validation with 50 iterations is selected for finding the best hyperparameters of the estimator and the model with selected features.

Following tuning the estimator, the overfitting of the selected estimator was checked using 5-fold cross-validation. Then, to identify the most important factors, the RFECV approach with tunned estimator was employed along with a standard scaler using pipe function for simultaneous scaling data and finding important features. Lastly, to determine the effectiveness of the dimensionality reduction, the model with selected features was compared with the full model using R-squared for the regression analysis and F1 score and Recall goodness of fit criteria for the classification analysis. The imbalanced datasets was the main challenge of classification analysis. In this study, since there are fewer broken pipes than non-broken, the model might not be able to learn from the training set and predict the broken pipes correctly. In these cases, although the model might perform excellent in general and have a high F1 score, it cannot predict the broken pipes accurately. Hence, to ensure the dimensionality reduction loses no information, the performance of the full model and selected model was compared in terms of recall score in addition to F1 score. Recall indicates the percentage of the actual positives (broken pipes) identified correctly.

For a more accurate comparison, prior to this step, the tuned hyperparameters of the model with selected features were determined, and the overfitting of the model was checked. It is noticeable that the RFECV analysis requires numerical targets and predictors. Hence, prior to the analysis, in addition to the categorical predictors, the values of the break status target attribute were converted to numerical using optimal scaling. The assigned values to the non-broken and broken pipe were 1 and 2, respectively. However, due to the nature of the XGBOOST classifier estimator, the target attribute should be binary. Hence, the converted values of 1 and 2 in the target attribute were replaced by 0 and 1, respectively.

4 Available dataset

This study is part of the project “Best Practices for Predicting WaterMain Breaks,” a collaboration between the National Water and Wastewater Benchmarking Initiative (NWWBI) and the Concordia University research group “UrbanLinks”. Thirteen utilities across Canada, including Barrie, Calgary, Region of Durham, Halifax, Kitchener, Region of Markham, Region of Waterloo, Saskatoon, St. John’s, Vancouver, Victoria, Waterloo, and Winnipeg, has shared their water main inventories and historical records of main breaks as separate spreadsheets or GIS shapefiles. The inventory file contains information on the characteristics of existing pipes in the system, and the break file lists the failure records of broken pipes. The list of available attributes for the two target attributes, the current rate of failure and break status, is provided in Table 4-1 and Table 4-2, respectively.

As the tables illustrate, overall, the provided data for this study can be categorized into five general groups of pipe physical characteristics, historical information, protection activities, environmental, and operational factors. Due to the lack of proper data collection framework, the available attributes were varied between each utility, yet some characteristics such as diameter, material, length, installation year, and failure year were consistently collected by all utilities. Also, the majority of the utilities record information such as lining status and lining material. Given that, a few information such as soil type, pressure, and casing material is recorded by a few utilities only.

Table 4-1 Available attributes provided by each utility for the current rate of failure

	Physical								Historical					Protection									Operational		Env
Utility	Diameter	Length	Joint type	Restrained	Material	Roughness	Dead end	Pipe Depth	Age	Failure month	Install month	Status	Previous rate of failure	Cathodic Protection stat	Cathodic Protection age	Lining status	Lining material	Lining age	Coating material	Casing material	Anode Type	Anode Status	Service type	Pressure	Soil type
Barrie	■	■		■	■		■	■	■	■	■	■	■	■						■		■	■		■
Calgary	■	■			■		■		■			■	■	■					■		■				
Durham	■	■			■			■	■	■	■	■	■	■	■	■	■	■							
Halifax	■	■			■				■	■	■		■			■	■								
Kitchener	■	■			■				■	■		■	■			■	■	■				■			
Markham	■	■			■				■	■	■		■	■	■	■									
Region of Waterloo	■	■			■				■	■		■	■			■	■	■							
Saskatoon	■	■	■		■				■				■			■	■	■							
St.John’s	■	■			■	■			■	■			■												
Vancouver	■	■			■			■	■	■			■				■		■				■		
Victoria	■	■			■				■	■			■			■	■							■	
Waterloo	■	■			■				■	■		■	■			■	■	■							
Winnipeg	■	■	■		■				■	■		■	■						■						

Table 4-2 Available attributes provided by each utility for break status target attribute

	Physical							Historical						Protection							Operational	
Utilities	Diameter	Length	Joint type	Restrained	Material	Roughness	Dead end	Age	Failure month	Install month	Status	Previous failure	Replaced Status	Cathodic Protection stat	Cathodic Protection age	Lining status	Lining material	Lining age	Coating material	Casing material	Service type	Pressure
Barrie	▨	▨		▨	▨			▨	▨		▨	▨		▨						▨	▨	
Calgary	▨	▨			▨		▨	▨				▨										
Durham	▨	▨			▨			▨	▨		▨	▨		▨	▨	▨	▨	▨				
Halifax	▨	▨			▨			▨	▨			▨				▨	▨					
Kitchener	▨	▨			▨			▨	▨			▨				▨	▨	▨				
Markham	▨	▨			▨			▨	▨			▨		▨	▨	▨		▨				
Region of Waterloo	▨	▨			▨			▨	▨		▨	▨				▨	▨	▨				
Saskatoon	▨	▨	▨		▨			▨			▨	▨	▨			▨	▨	▨				
St.John's	▨	▨			▨	▨		▨	▨			▨										
Vancouver	▨	▨			▨			▨	▨								▨		▨		▨	
Victoria	▨	▨			▨			▨	▨			▨				▨	▨					▨
Waterloo	▨	▨			▨			▨	▨	▨		▨				▨	▨	▨			▨	
Winnipeg	▨	▨			▨			▨	▨		▨	▨							▨			

It is noticeable that while the utilities directly recorded the majority of the attributes, some others, including age, failure month, install month, the previous rate of failure, protection age, and lining age, were extracted from the available information to evaluate their impacts on failures. For instance, lining age for the lined pipes is the difference between pipe failure year, current available failure year for non-broken pipes, and lining installation year. For pipes without lining, the lining age is considered as 0. Not all utilities have collected lining installation year. Figure 4-1compares the average lining age for broken and inventory pipes for the utilities whose recorded this attribute.

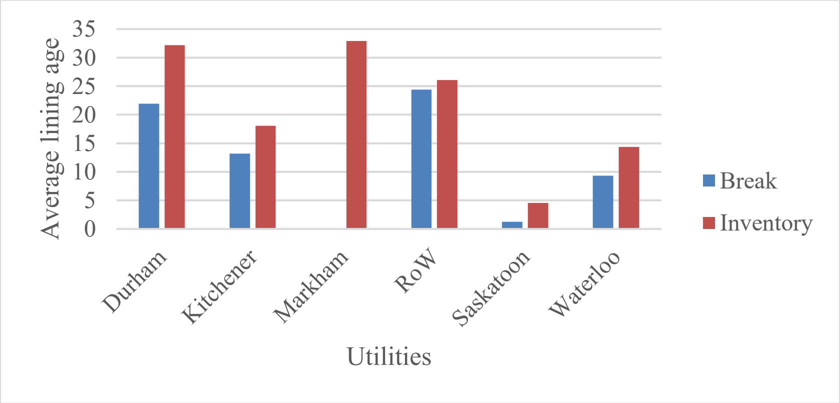


Figure 4-1 Average lining age of broken and inventory pipes for the utilities with available lining age

Figure 4-1 indicates the average age of lining for the broken pipes in the system is around 14 years and for inventory pipes is approximately 32 years. Meaning that majority of the lined pipes in the system last more than 14 years after lining.

As mentioned earlier in the literature, age, as an indicator of many other factors such as external load, corrosion, etc., plays a vital role in various deterioration models. The average age of the broken and inventory pipes in the analyzed utilities in this study is around 46 years and 36 years, respectively. Figure 4-2 presents information on the average age of each utility's broken pipes and inventory pipes. Evaluating the distribution of breaks during different months of the year indicates, majority of the pipe failure has occurred during cold winter months, i.e., January and February as presented in Figure 4-3.

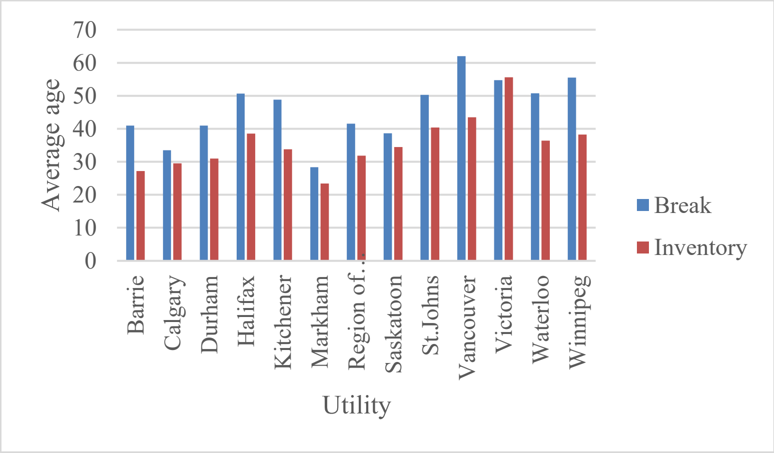


Figure 4-2 Average age of pipes for each utility

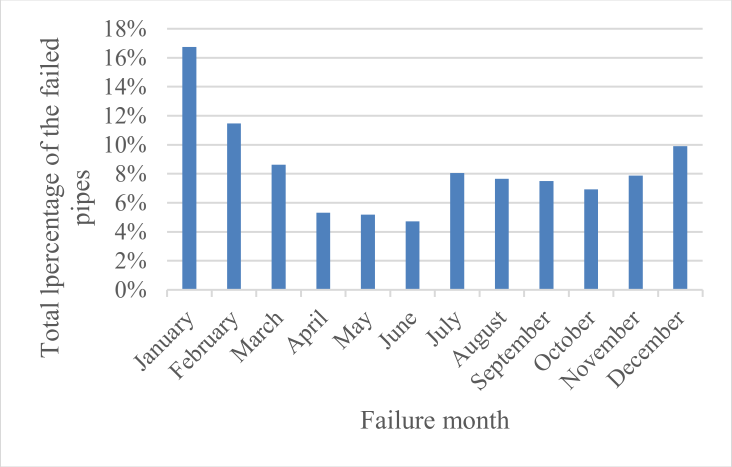


Figure 4-3 Total percentage of the failed pipes in each month

The general characteristics of the available pipes in the inventory and broken pipes for all utilities are provided in Table 4-3. It is noticeable the mentioned values are related to the clean datasets prepared for targets of the rate of failure and break status. According to the Table 4-3, this study analyzed around 24835 km of inventory pipes and 3167 km of broken pipes with an average of 4771 breaks. The available break year data varies between different utilities, indicating utilities started collecting data on broken pipes in different time frames. More specifically, Calgary, Saskatoon, and Winnipeg have the most historic break data, while break information i Vancouver and Waterloo is the most recent.

Table 4-3 General characteristics of pipes in each utility

Attribute	Length (Inventory) KM	Length (Break) KM	Break Decades Available							Number of breaks per KM
			50	60	70	80	90	00	10	
Barrie	897	61			■	■	■	■	■	15.1
Calgary	6811	1048	■	■	■	■	■	■	■	15.1
Durham	3183	440			■	■	■	■	■	13.3
Halifax	2710	384			■	■	■	■	■	14.9
Kitchener	12	2				■	■	■	■	956
Markham	1501	147			■	■	■	■	■	15.7
Region of Waterloo	392	18				■	■	■	■	10.2
Saskatoon	1363	238	■	■	■	■	■	■	■	27.4
St.John's	694	96				■	■	■	■	15.4
Vancouver	1577	29						■	■	25.5
Victoria	351	74				■	■	■	■	9.3
Waterloo	481	53						■	■	13.4
Winnipeg	4862	576	■	■	■	■	■	■	■	33.2

Overall, the initial analysis of the data for all utilities indicates that by improving the manufacturing process of different materials over the years, the percentage of pipe material has changed, as demonstrated in Figure 4-4. According to the mentioned figure, the cast iron pipes were commonly used during 1900-1960. However, this pattern shifted in 1970 when the ductile iron pipes were widely used in the system and PVC afterward in 1980.

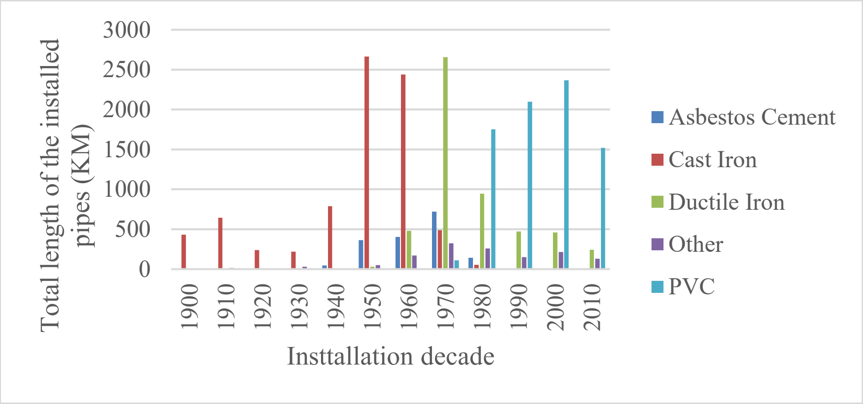


Figure 4-4 Total lengths of the pipes installed in each decade

Since cast iron pipes are the oldest pipes in the system in this study, they accounted for most of the failures in all utilities. The information related to the material of the failed pipes and inventory pipes of all utilities is summarized in Figure 4-5Figure 4-5 and Figure 4-6 respectively.

Overall, in this analysis, more than half of the failed pipes were cast-iron pipes according to Figure 4-5. Also, around 40% of the inventory pipes were made from PVC material followed by cast iron and Ductile iron as presented in Figure 4-6. As presented in Figure 4-4, cast-iron pipes are the oldest pipes in the system, and their higher failures compared with other materials can be explained accordingly.

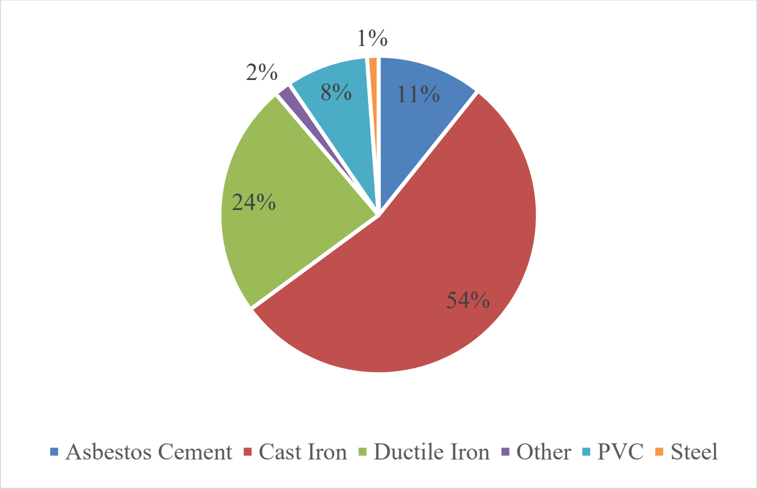


Figure 4-5 Total lengths of the failed material

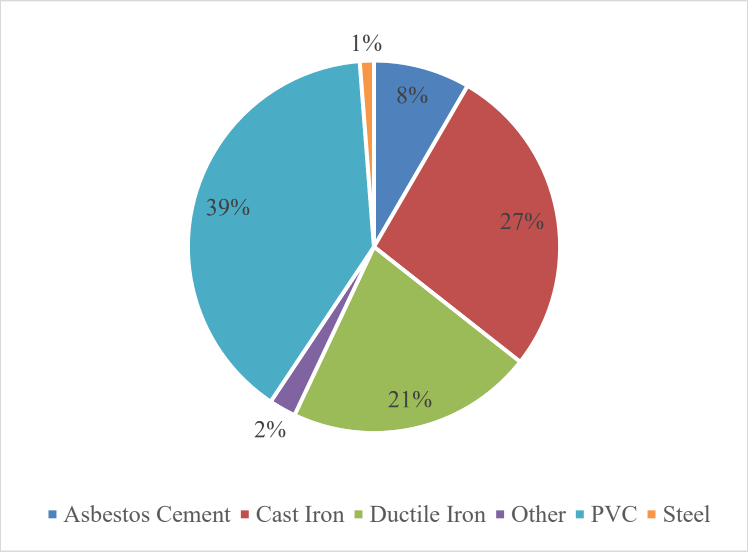


Figure 4-6 Total percentage of the inventory pipes material

It is noticeable that the given Figure 4-5 and Figure 4-6 are related to all utilities, and the detail of material in each utility might be different depending on the available break year. The more detailed material for each utility and protection properties for each utility are provided in Table 4-4 and Table 4-5 for break and inventory files, respectively.

As presented in Table 4-4, cast iron is the predominant broken pipe material in all utilities except Markham and the Region of Waterloo, where ductile iron is the primary failed material. Also, cast-iron pipes that were failed on average at 55 years are the oldest failed pipes in all utilities, and PVC pipes with an average age of 16 years are the youngest failed pipes. According to the same table, Markham is the only utility widely employed corrosion protection activities in their water network.

Table 4-4 Broken pipes detailed material and protection characteristics

Utility	% CI pipes	% DI pipes	%PVC Pipes	Average age of CI	Average age of DI	Average age of PVC	% Lined pipes	% Protected pipes
Barrie	58	30	7	48	31	13	-	3
Calgary	49	36	9	45	25	3	-	1
Durham	38	33	18	51	33	33	21	32
Halifax	74	20	3	55	26	18	33	-
Kitchener	70	27	2	55	37	12.	1	-

Utility	% CI pipes	% DI pipes	%PVC Pipes	Average age of CI	Average age of DI	Average age of PVC	% Lined pipes	% Protected pipes
Markham	32	53	11	37	26	13	52	72
RoW	36	40	8	62	34	16	12	-
Saskatoon	37	0.2	10	61	37	5	2	-
St.John's	82	15	1	56	27	3	-	-
Vancouver	88	4	0.1	64	33	45	28	-
Victoria	72	23	3	64	27	18	5	-
Waterloo	81	14	5	54	42	22	23	-
Winnipeg	65	2	4	67	32	10	-	-

While PVC is the current primary material in most cities in this analysis, the existing pipes in Halifax, St. John's, Vancouver, and Victoria are either ductile iron or cast iron, as illustrated in Table 4-5. According to the same table, the average age of the PVC pipes in the system is around 18 years, indicating the utilities recently employed this material. Also, the average of the current cast-iron pipes in the system is 60 years, given that this material still constitutes a large portion of the existing pipes in the system. These very old CI pipes are the main cause of the problem in different utilities as presented in Figure 4-5. The more detailed data summery is provided in ()

Table 4-5 Inventory pipes detailed material and protection characteristics

Utility	% CI pipes	% DI pipes	%PVC Pipes	Average age of CI	Average age of DI	Average age of PVC	% Lined pipes	% Protected pipes
Barrie	16	25	52	51	32	17	-	4.3
Calgary	21	23	47	48	34	18	-	-
Durham	17	20	54	55	37	21	13	13
Halifax	44	48	2	56	24	21	44	-
Kitchener	24	35	36	61	38	13	0.3	-
Markham	9	14	71	32	39	19	11.2	27
RoW	10	32	69	69	36	17	25	-
Saskatoon	19	0.2	68	68	46	17	0.9	-
St.John's	43	44	12	64	27	6.2	-	-
Vancouver	43	54	0.09	68	24	22	47	-
Victoria	49	37	7	78	35	28.2	5	-
Waterloo	31	15	53	59	43	21	11	-
Winnipeg	25	1	55	70	35	22	-	-

5 Results

This section provides a general description of correlation analysis results, and a detailed explanation of dimensionality reduction approaches results for the two target attributes. This section in the first part describes the results of correlation analysis for both targets. Then in the following sections, the results of each dimensionality reduction approach and a general comparison of each within different cities are provided.

Overall, four dimensionality reduction approaches have been employed to identify the most important factors affecting the current watermain failures across thirteen Canadian utilities. These approaches are FAMD, CATPCA, RF-RFECV, and XGBOOST-RFECV.

A more detailed explanation of the results of each analysis for each target attribute is provided in the following sections.

5.1 Correlation analysis

This section provides an overview of the results of correlation analysis. As mentioned earlier, although the utilities collected different subsets of data, a few attributes were consistently collected by all of them. The correlation coefficients are the values ranging between -1 and 1 which the absolute values indicate the significance of relations between two attributes. The larger absolute values indicate the stronger relationships. In this analysis, the negative values indicate the inverse relationship between variables, and the positive values represent a direct relationship.

The common attributes for the current rate of failure target are material, diameter, length, age, and previous rate of failure. As presented in Figure 5-1, these attributes are neither correlated with each other nor with the target for the current rate of failure. Among these attributes, the previous rate of failure has the most significant correlation with the target, and still, it is not highly correlated with it.

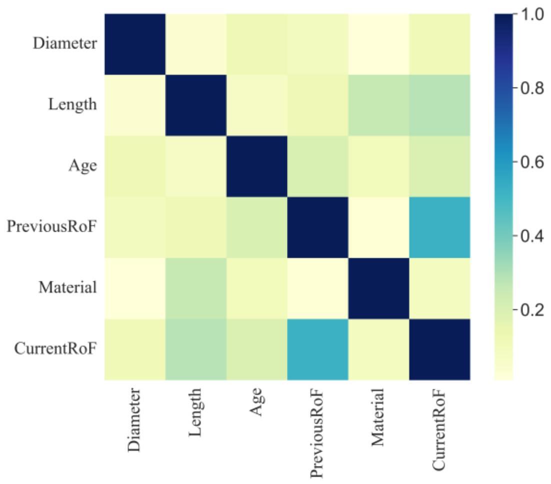


Figure 5-1 Correlation analysis - Common attributes & Current rate of failure

In the break status analysis, the material, diameter, length, and age are common factors. The results of correlation analysis between these attributes and the target are provided in Figure 5-2. Similar to the previous one, the attributes are neither highly correlated with each other nor with the target in this analysis as well. Among these common attributes, material and the length of the pipes has the strongest association with the target. However, this relationship is not strong enough so that the attributes cannot be considered highly correlated.

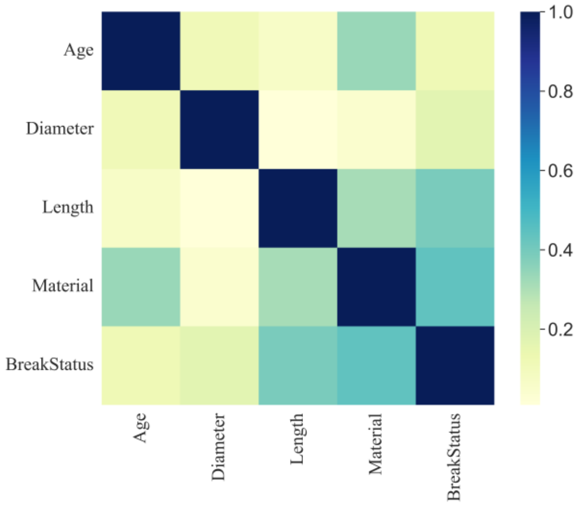


Figure 5-2 Correlation analysis - Common attributes & Break status

Also, in order to realize the relationship between different attributes and the targets, a separate correlation analysis was performed on data of each utility. The correlation coefficients for each attribute are then employed to create boxplots presented in Figure 5-3 and Figure 5-4 for the current rate of failure and break status, respectively. The boxplots indicate the range of correlation coefficients between each attribute and the target. The number of values in each box plot depends on how many utilities record that information. Accordingly, for the attributes collected by one utility only, the values are represented as a line instead of a box.

As presented in Figure 5-3, in the current rate of failure analysis, except for pipe length that is highly correlated with the target, the remaining attributes are not correlated with the target. As mentioned earlier, the current rate of failure indicates the total number of failures each meter of pipe has experienced in its latest available failure year. Hence, the strong correlation between length and current rate of failure is expected. Beside length, the previous rate of failure is somehow correlated with the target.

Given that, in the next section, results of CATPCA indicate although these attributes are not highly correlated with the target, some of them are significantly important in predicting the current rate of failure.

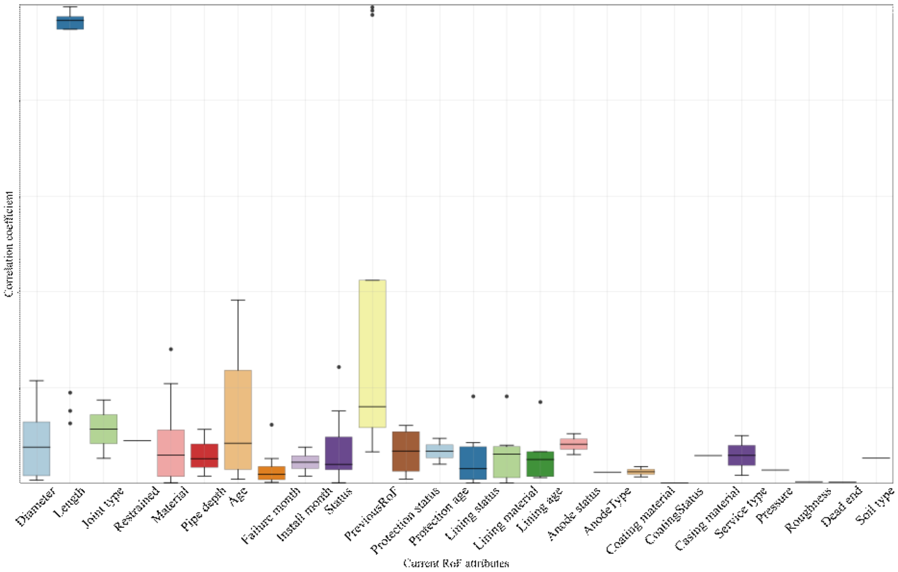


Figure 5-3 Correlation coefficients - Current rate of failure

The results of correlation analysis for the break status target are provided in Figure 5-4. According to the results, in this analysis, the correlation coefficients are higher than the current failure rate analysis and the previous number of failures is the only highly correlated factor with the target. Many factors can prevent the failure, however, from the literature, we know once a pipe breaks, it tends to fail in other locations. This confirms the dependency of previous rate of failure to broken pipes and explains why in general the correlation coefficients in break status analysis are higher than the current failure rate.

In Figure 5-4 besides the previous number of failures, comparing the correlation coefficients indicates a higher correlation coefficient between pipe material and the target than other attributes in this analysis. As mentioned in the literature, different materials have different expected life services, confirming the stronger correlation between break status and material. Also, while previous studies had commonly used the age of the pipes in watermain failure predictions, this analysis's results indicate that protection information, i.e., protection age and protection status, are more associated with the target than age. This highlights that the deterioration of pipes can be highly affected by protection activities. More specifically, the corrosion protection information such as the material used for it and the year it performed on the pipe are more important than age in predicting watermain failure.

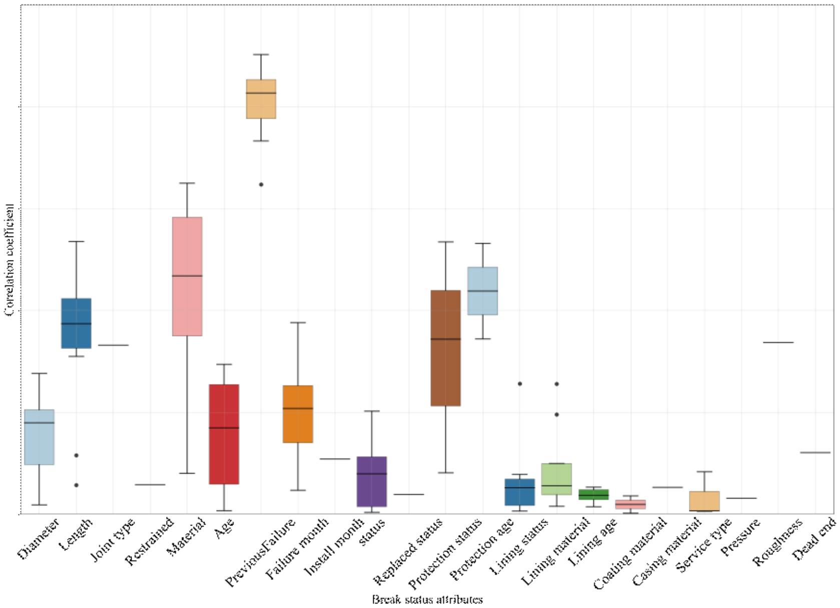


Figure 5-4 Correlation coefficients - Break status

5.2 Current rate of failure analysis

The current rate of failure is a numerical target attribute which indicates the number of failures that occurred in the latest available failure year per length. In order to identify the most important factors affecting this target, three types of analysis were performed: FAMD, CATPCA, and RFECV. Results of each method are explained in more detail in the following sections.

5.2.1 Factor Analysis of Mixed Data (FAMD)

FAMD estimates the contribution of each attribute to the target. Figure 5-5 represents the results of the FAMD analysis for the city of Winnipeg, as an example. The red line in the figure represents the average contribution. For each utility, the most important factors are the ones with the contribution greater than the average. The average contribution is the uniform contribution of variables and is equivalent to 1 over the total number of variables.

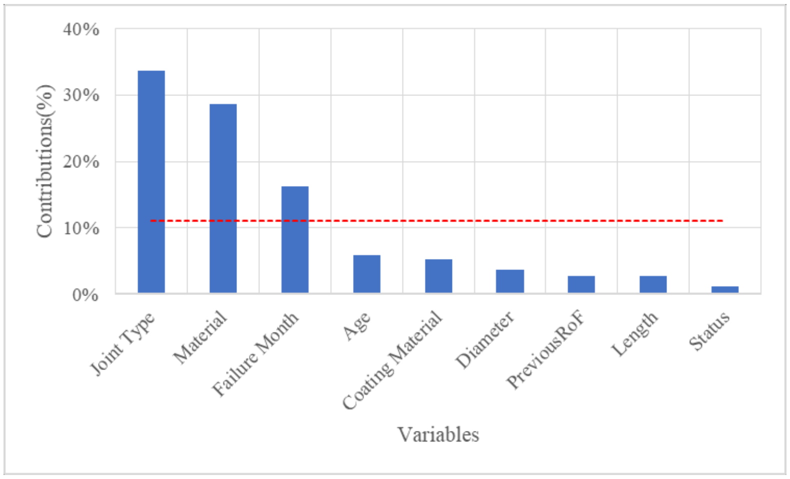


Figure 5-5 Overall contribution of variables in principal components - city of Winnipeg – Current rate of failure

The results of FAMD analysis for the current failure rate are provided in Table 5-1. In the provided table, the available attributes and the most important attributes are highlighted in yellow and orange, respectively. Also, the last line of the table indicates the cut-off level for selecting the important features. Comparing the results over different cities indicates that FAMD rates categorical data more important than numerical ones. More specifically, the results consistently identified material (4-10 categories), failure month (12 categories), and where available soil type (33 categories), joint type (7 categories), and casing material (5 categories) as the most important factors affecting the pipes' failure rate in all utilities. The remaining categorical variables were also selected among the key factors occasionally depending on their number of categories compared with the other categorical variables. For instance, install month in Barrie has only two categories that are negligible compared with the other categorical variables in the city. Also, the lining material is among the most important factors for a few utilities recording it. Given that, none of the numerical variables were selected as key factors in predicting the current failure rate according to FAMD.

As mentioned earlier in the literature, a linear relationship between the target and the predictors is a key assumption in FAMD, whereas the results of correlation analysis indicate no linear relationship in this study.

Table 5-1 FAMD results – Current RoF (Orange - important, yellow – not important, blank – not available)

	Attributes	**Barrie**	**Calgary**	**Durham**	**Halifax**	**Kitchener**	**Markham**	**RoW**	**Saskatoon**	**St.Johns**	**Vancouver**	**Victoria**	**Waterloo**	**Winnipeg**
Physical	**Joint type**								**32**					**34**
	Diameter	3	6	4	5	4	4	5	6	8	6	5	4	4
	Material	**18**	**51**	**15**	**20**	**25**	**24**	**23**	**35**	**35**	**26**	**32**	**13**	**29**
	Length	2	3	3	3	4	3	5	3	6	3	3	4	3
	Restrained	1												
	Roughness									7				
	Dead-end		4											
	Pipe Depth	2					4				4			
Historical	**Failure Month**	**17**		**21**	**21**	**29**	**23**	**27**		**30**	**28**	**27**	**42**	**16**
	Install Month	4		**22**	**27**		**20**							
	Status	2	6	2		3		4					4	1
	Age	3	6	5	5	5	4	5	6	7	5	6	4	6
	PreviousRoF	3	3	3	5	4	3	4	2	6		3	4	3
Protection	**Casing Material**	**11**												
	Lining Mateial			7	8	**12**		6	7		**13**	12	**11**	
	Lining Status			5	7	6	5	6	7			7	7	
	Lining Age			4		6		6	3				6	
	Protection Status	2	6	5			5							
	Protection Age			5			5							
	Coating Material										8			5
	Coating Status		5											
	Anode type		9											
	Anode status	2				3								
Operational	**Service type**	3									7			
	Pressure											4		
Env	**SoilType**	**28**												
Contribution cut-off level		7%	10%	8%	11%	9%	9%	10%	11%	14%	11%	11%	10%	11%

5.2.2 Categorical Principal Component Analysis (CATPCA)

CATPCA estimates the contribution of numerical and categorical variables to predicting the target. Similar to the FAMD method, the cut-off level for identifying the most important factors is determined according to the number of attributes for each utility. The overall results of the CATPCA analysis are presented in Table 5-2. In the table the available attributes and the most important attributes are highlighted in yellow and orange, respectively. Also, the last line of the table indicates the cut-off level for selecting the important features.

Table 5-2 CATPCA results - Current RoF (Orange - important, yellow – not important, blank – not available)

	Attributes	Barrie	Calgary	Durham	Halifax	Kitchener	Markham	RoW	Saskatoon	St. John's	Vancouver	Victoria	Waterloo	Winnipeg
Physical	Joint type								10					9
	Diameter	7	11	6	14	10	10	6	10	14	10	9	12	12
	Material	7	10	7	8	11	6	10	11	13	11	7	7	8
	Length	8	13	6	13	13	10	6	11	17	14	10	9	11
	Restrained	6												
	Roughness									12				
	Dead-end		10											
	Pipe Depth	5					7				12			
Historical	Failure Month	7		8	13	8	12	10		15	13	10	11	11
	Install Month	7		8	7		8							
	Status	7	11	8		10		11					11	12
	Age	7	10	7	10	13	7	11	10	12	8	10	7	7
	PreviousRoF	8	6	8	9	8	9	10	12	16		15	8	18
Protection	Casing Material	5												
	Lining Mateial			9	13	8		12	12		9	14	12	
	Lining Status			9	13	7	9	12	12			14	12	
	Lining Age			8		6		12	12				11	
	Protection Status	7	9	9			11							
	Protection Age			9			11							
	Coating Material										11			12
	Coating Status		10											
	Anode type		10											
	Anode status	7				6								
Operational	Service type	5									11			
	Pressure											10		
Env	SoilType	7												
Contribution cut-off level		7%	10%	8%	11%	9%	9%	10%	11%	14%	11%	11%	10%	11%

In general, the results of the CATPCA analysis indicate the importance of protection activities in predicting the current rate of failure. More specifically, the results indicate lining information is consistently important in all utilities except Kitchener. According to the results, it is recommended that utilities with more than 30% of protected pipes, i.e., Durham and Markham, track information related to protection properties such as when the protection is performed on the pipes. Other protection activities include coating, casing, and anodic protection. These attributes were collected for a few utilities, and although the analysis results have rated coating properties and anode information among the essential factors, their actual impact requires further investigation. The results highlight the importance of the historical information and physical characteristics for utilities without broad records of protection activities, e.g., Barrie, in predicting the current failure rate. In this analysis, status and where available soil type were also consistently identified as important factors affecting the current failure rate. Comparing the results of CATPCA and FAMD indicates while in FAMD, the important factors were selected from the categorical variables, a combination of categorical and numerical variables are important herein. As mentioned earlier, although the CATPCA can handle the linear and non-linear relationship between the variables and the target, FAMD, which is based on PCA, and MCA is capable of linear relationships only. Hence, the differences between the two approaches in selecting the most important features sound logical.

5.2.3 Recursive Feature Elimination with Cross-Validation (RFECV)

RFECV identifies the most contributing factors in predicting the target through a recursive process using a supervised learning algorithm as an estimator. In this method, initially, the highly correlated attributes were removed from the analysis. The significant level of 0.8 was considered for determining the highly correlated attribute as suggested by Jun et.al. (2020) on studying the different factors affecting steel water-transmission main rate of failure in Korea. The lining information and protection information was the only highly correlated attributes. Hence, where available, lining age and protection age were kept as representers of lining and protection. For the cities without lining age, lining material was kept only. The correlation matrices for each city are provided in Appendix B.

For this study, Random Forest and XGBOOST were selected as estimators of RFECV analysis. The overall results of the RF-RFECV and the XGBOOST-RFECV are provided in Table 5-3 and Table 5-5, respectively. The total number of predictors for each utility is also provided in the same table. In this analysis for both estimators, the length of the pipes was revealed as a vital factor affecting the current rate of failure. In order to evaluate the impacts of length on this target, intentionally, the length of the pipes was removed from the data, and a separate analysis was performed accordingly. Since the performance of the models without length was highly reduced compared with the initial models, the length of the pipes was kept as a predictor in the analysis.

The more detailed results of each analysis will be explained in 5.2.3.1 and 5.2.3.2 for random forest and XGBOOST, respectively. It can also be shown how the number of predictors can affect a model’s performance. For instance, Figure 5-6 indicates how the number of attributes affects the performance of the Random Forest model in predicting the current rate of failure in Calgary.

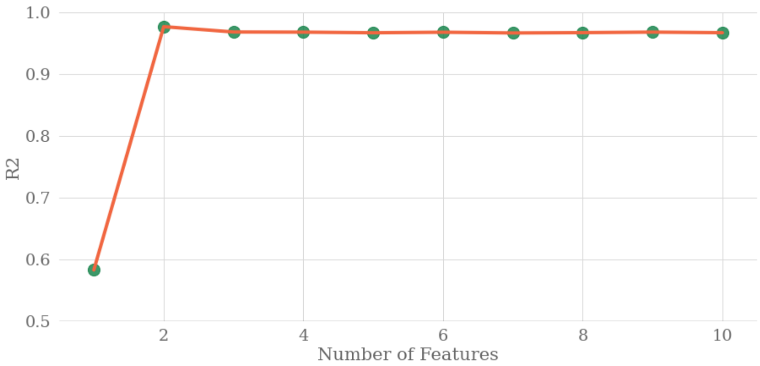


Figure 5-6 Number of attributes vs. performance of model - city of Calgary

5.2.3.1RF-RFECV

The overall results of RF-RFECV analysis are provided in Table 5-3. In the table, the most important features are highlighted in orange and are assigned a weight by the random forest estimator. These weights sum one and indicate the contribution of each variable in predicting the targets. Accordingly, the higher weights, the more contributing factors. Comparing the selected features with the available predictors for each city indicates this approach significantly reduces the number of features, and still, the models with the selected features perform equally or better than the full models. Overall, the analysis results identified physical, historical, and, where available environmental factors as the most important factors affecting pipes' current rate of failure. In this analysis, unlike CATPCA, protection activities do not play an essential role in predicting the current failure rate. In the majority of the cities, the estimator selected the pipe's length alone or along with a few other common physical or historical factors to predict the current failure rate more precisely. It is noticeable that in the cities the estimator did not select the length of the pipe as an important factor, age and previous rate of failure are the two critical elements that were chosen instead. According to the results, length, age, and previous rate of failure are the most contributing factors in predicting the current rate of failure.

Table 5-3 RF- RFECV weights and results - Current RoF (Orange - important, yellow – not important, blank – not available)

	Attributes	Barrie	Calgary	Durham	Halifax	Kitchener	Markham	RoW	Saskatoon	St. John's	Vancouver	Victoria	Waterloo	Winnipeg
Physical	Joint type								🟨					🟨
	Diameter	🟨	🟨	0.002	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
	Material	🟨	🟨	0.01	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	0.002
	Length	0.99	🟨	0.96	1	0.93	1	1	1	1	1	🟨	1	0.16
	Restrained	🟨												
	Roughness									🟨				
	Deadend		🟨											
	Pipe depth	🟨					🟨				🟨			
Historical	Failure month	0.001		0.01	🟨	🟨	🟨	🟨		🟨	🟨	🟨	🟨	0.004
	Install month	🟨		0.00	🟨		🟨							
	Status	🟨	🟨	0.0006		🟨		🟨					🟨	🟨
	Age	0.001	0.81	0.01	🟨	0.07	🟨	🟨	🟨	🟨	🟨	0.12	🟨	0.040
	Previous RoF	🟨	0.18	0.001	🟨	🟨	🟨	🟨	🟨	🟨		0.88	🟨	0.79
Protection	Casing material	🟨												
	Lining material			🟨	🟨	🟨		🟨	🟨		🟨	🟨	🟨	
	Lining status			🟨	🟨	🟨	🟨	🟨	🟨			🟨	🟨	
	Lining age			🟨		🟨		🟨	🟨				🟨	
	Protection status	🟨	🟨	0.0005			🟨							
	Protection age			🟨			🟨							
	Coating material										🟨			🟨
	Coating status		🟨											
	Anode type		🟨											
	Anode status	🟨				🟨								
Operational	Service type	🟨									🟨			
	Pressure											🟨		
Env	Soil type	0.002												
Total attributes		14	10	13	9	11	11	10	9	7	9	9	10	9
Full model r2		87.9	97.5	98	96.3	90.9	96.2	96.2	96.7	99.4	98.7	96.9	99.53	99.15
Final model r2		94.15	98.7	98.2	97.4	92.5	96.6	96.6	97.8	99.6	99.1	98.6	99.56	99.16

Since the current rate of failures was calculated using age and due to the large contribution of this attribute in prediction the current rate of failure as well as the poor performance of the model without it, using random forest, the available attributes were employed to predict the number of failures, i.e., an independent target, instead. The performance of the model is provided in Table 5-4. According to the results, the performance of models for predicting the current number of breaks is poor in all utilities. Hence, prediction of number of breaks alone are not accurate.

Table 5-4 Performance of current number of breaks vs current rate of failure random forest prediction models

Utilities	Full model Number of breaks r2	Full model rate of failure r2
Barrie	-16	87.9
Calgary	10.6	97.5
Durham	1.7	98
Halifax	3.1	96.3
Kitchener	-6.9	90.9
Markham	-2.5	96.2
RoW	7.4	96.2
Saskatoon	5.1	96.7
St. John's	-4.2	99.4
Vancouver	-3.6	98.7
Victoria	-1.8	96.9
Waterloo	-2	99.53
Winnipeg	2.9	99.15

5.2.3.2XGBOOST-RFECV

The overall results of the XGBOOST-RFECV analysis are provided in Table 5-5. Similar to RF-RFECV, this approach significantly reduces the number of predictors, and the models with selected features perform better or equally compared with the full model. The pattern of the final result in this approach is almost similar to RF – RFECV. The overall results of the analysis indicate this method widely selects physical factors and historical information as the two groups of important factors and do not consider the remaining categories as influential factors for predicting the current rate of failure. More specifically, the length of the pipes was consistently identified as an important factor in all utilities except Victoria where previous rate of failure is the only contributing factor. Overall, in this analysis, in most of the utilities, the length of the pipe alone or with a few other attributes, including diameter, material, age, and previous rate of failure, are the essential factors in predicting the current rate of failure. In this analysis, among the selected factors, length and the previous rate of failure are the most contributing factors in all utilities.

Comparing the results of RF and XGBOOST models based on the resulting R^2 indicates that the random forest approach generally predicts the current failure rate more accurately.

Table 5-5 XGBOOST – RFECV weights and results – Current RoF (Orange - important, yellow – not important, blank – not available)

Attributes		Barrie	Calgary	Durham	Halifax	Kitchener	Markham	RoW	Saskatoon	St. John's	Vancouver	Victoria	Waterloo	Winnipeg
Physical	Joint type								🟨					🟨
	Diameter	🟨	🟨	0.02	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
	Material	🟨	🟨	0.1	🟨	🟨	🟨	🟨	🟨	🟨	0.02	🟨	🟨	🟨
	Length	1	0.12	0.85	1	0.98	1	1	0.26	🟨	0.96	🟨	0.12	0.55
	Restrained	🟨												
	Roughness									🟨				
	Deadend		🟨											
	Pipe depth	🟨					🟨				🟨			
Historical	Failure month	🟨		🟨	🟨	🟨	🟨	🟨		🟨	0.006	🟨	🟨	🟨
	Install month	🟨		0.02	🟨		🟨							
	Status	🟨	🟨	🟨		🟨		🟨					🟨	🟨
	Age	🟨	0.15	🟨	🟨	🟨	🟨	🟨	0.13	1	0.008	🟨	0.15	0.01
	Previous RoF	🟨	0.72	🟨	🟨	🟨	🟨	🟨	0.6	🟨		1	0.72	0.4
Protection	Casing material	🟨												
	Lining material			🟨	🟨	🟨		🟨	🟨		🟨	🟨	🟨	
	Lining status			🟨	🟨	🟨	🟨	🟨	🟨			🟨	🟨	
	Lining age			🟨		🟨		🟨	🟨				🟨	
	Protection status	🟨	🟨	🟨			🟨							
	Protection age			🟨			🟨							
	Coating material										🟨			🟨
	Coating status		🟨											
	Anode type		🟨											
	Anode status	🟨				0.002								
Operational	Service type	🟨									🟨			
	Pressure											🟨		
Env	Soil type	🟨												
attributes		14	10	13	9	11	11	10	9	7	9	9	10	9
model r2		98.3	93	96.7	88	82	93.2	91	94	96	98	57	93	95.1
model r2		98.3	93.3	96.8	96	83	93.2	92	94	97	99	57	93.3	95.7

5.3 Break status

5.3.1 Factor Analysis of Mixed Data (FAMD)

The overall results of the analysis for the break status target are provided in Table 5-6Table 5-1. In the table, the last line indicates the cut-off level for selecting the important features. The available and most important attributes are highlighted in yellow and orange in the table, respectively. As mentioned earlier, the results of this approach are biased since this approach is mainly rated categorical attributes more important than numerical ones, and usually, categorical variables with more categories are more important. Hence, only the categorical predictors are among the most important attributes. More specifically, in this analysis, material (7-12 categories) and where available joint type (7 categories), install month (12 categories), casing material (7 categories), and coating material (6-9 categories) were consistently selected as the most important factors affecting the deterioration of pipes.

Table 5-6 FAMD results- Break Status (Orange - important, yellow – not important, blank – not available)

	Attributes	Barrie	Calgary	Durham	Halifax	Kitchener	Markham	RoW	Saskatoon	St.Johns	Vancouver	Victoria	Waterloo	Winnipeg
Physical	**Joint type**								**24**					
	Diameter	6	11	5	7	6	5	5	4	7	7	5	3	6
	Material	**33**	**56**	**22**	**41**	**31**	**41**	**32**	**35**	**39**	**26**	**34**	**16**	**43**
	Length	3	8	3	5	4	4	3	3	6	2	4	3	6
	Restrained	3												
	Roughness									7				
	Dead-end		7											
Historical	**Failure Month**	**17**		**30**	**14**	**19**	**21**	**26**		**25**	7	**19**	**24**	8
	Install Month												**18**	
	Status	4		2				4	1					3
	Age	5	11	4	6	7	4	6	5	9	7	6	5	8
	Previous Failure	4	8	4	6	5	4	4	4	6		5	4	7
	Replaced Status								5					
Protection	**Casing Material**	**14**												
	Lining Mateial			**9**	**14**	**15**		7	8		**20**	**16**	**12**	
	Lining Status			5	8	7	6	7	6			7	6	
	Lining Age			5		7	5	6	5				6	
	Protection Status	4		5			6							
	Protection Age			5										
	Coating Material										**18**			**18**
Operational	**Service type**	6									**14**		1	
	HGL											3		
Contribution cut-off level		9%	17%	8%	13%	11%	11%	10%	9%	14%	13%	11%	9%	13%

5.3.2 Categorical Principal Component Analysis (CATPCA)

The overall results of the CATPCA analysis for break status are provided in Table 5-7. The last row of the table indicates the cut-off level for identifying the most important factors. Attributes with the contribution of greater or equal than that cut-off are the important factors that are highlighted in orange in the table.

The overall results of CATPCA analysis for break status target highlight the significant impacts of protection activities on pipes' probability of failure. Interestingly, in this analysis, not only the status of protection was important, but also the material used for protection, i.e., lining material and age of that protection, did impact the probability of failure. Hence, besides corrosion protection status, how and when that protection is being put in place impacts deterioration. Among all protection activities information, coating material and casing status are only available for a few cities, and their actual impact on deterioration requires further investigation. The results also consistently identified the most common physical factors, i.e., material, diameter, and length, for the utilities without comprehensive protection activities data as presented for Barrie and Calgary. A few attributes were collected by one or two cities in this analysis. Although some of these attributes, including install month and pressure were identified as important factors, their actual impacts require further investigation.

Table 5-7 CATPCA results - Break Status (Orange - important, yellow – not important, blank – not available)

	Attributes	Barrie	Calgary	Durham	Halifax	Kitchener	Markham	RoW	Saskatoon	St. John's	Vancouver	Victoria	Waterloo	Winnipeg
Physical	**Joint type**								8					
	Diameter	**11**	**20**	8	10	**13**	**11**	**12**	**12**	**17**	12	10	3	13
	Material	**11**	**25**	**9**	**14**	12	10	10	**12**	14	12	**16**	**10**	12
	Length	**12**	**24**	5	**14**	**14**	9	**11**	**10**	**23**	11	8	**11**	**15**
	Restrained	5												
	Roughness									12				
	Dead-end		10											
Historical	**Failure Month**	**12**		**10**	**17**	**14**	**12**	10		**17**	**15**	**14**	**9**	**15**
	Install Month												**12**	
	Status	**11**		**9**				**11**	4					**16**
	Age	9	**21**	8	**14**	11	9	**11**	**12**	16	11	10	**10**	**15**
	Replaced Status								**10**					
Protection	**Casing Material**	8												
	Lining Mateial			**10**	**16**	**13**		**12**	**11**		**13**	**14**	**11**	
	Lining Status			**10**	**16**	12	**14**	**12**	**11**			**14**	**11**	
	Lining Age			**10**		12	**14**	**11**	9				**11**	
	Cathodic Protection Status	**12**		**10**			**11**							
	Cathodic Protection Age			**10**			**11**							
	Coating Material										**15**			**15**
Operational	**Service type**	**10**									11		**12**	
	Pressure											**14**		
	Contribution cut-off level	10%	20%	9%	14%	13%	11%	11%	10%	17%	13%	13%	10%	14%

5.3.3 Recursive Feature Elimination with Cross-Validation (RFECV)

This target consists of Yes and No values corresponding to the failed and non-failed pipes, respectively. These values were converted to numerical ones using optimal scaling. Accordingly, for the random forest estimator, the assigned values to the non-broken and broken pipe were 1 and 2, respectively. For the XGBOOST estimator however, these values were replaced by 0 and 1 instead as explained in the methodology. The overall results of RF-RFECV and XGBOOST-RFECV are presented in Table 5-8 and Table 5-10, respectively. In the tables, to ensure the dimensionality reduction loses no information, the performance of the full model and selected model was compared in terms of recall score in addition to F1 score. The following paragraphs describe the results of the RFECV approach with Random Forest and XGBOOST estimators.

5.3.3.1RF-RFECV

The overall results of RF-RFECV analysis are provided in Table 5-8. Comparing the selected features with the available predictors for each city indicates this approach significantly reduces the number of features, and still, the models with the selected features perform equally or slightly better than the full models in terms of F1 and recall scores. The high values of F1 and recall scores indicate the model's ability with selected features to accurately predict broken and non-broken pipes. Compared with the current rate of failure analysis, more features are selected by the random forest estimator in this analysis. As mentioned earlier, many factors can prevent failure, i.e. can affect pipe break status. Hence, selection of more factors for this target compared with the current rate of failure analysis was expected. Overall, the analysis has rated the physical and historical factors as the most important factors affecting deterioration. More specifically, the estimator consistently identified material, length, and age as the most important attributes affecting the deterioration. Out of the selected features, the mentioned three attributes are the most contributing factor in this analysis. Unlike CATPCA, the protection activity predictors do not affect the deterioration as high as physical and historical variables. According to the results, in the protection activities group, cathodic protection age and where available casing material were consistently highlighted as the important factors. The results also have rated joint type, pressure, roughness, and deadend among the most important features. However, as mentioned earlier, these attributes are available for a few cities only, and their actual impact on deterioration requires further investigation.

Table 5-8 RF-RFECV weights and results - Break Status (Orange - important, yellow – not important, blank – not available)

Attributes		Barrie	Calgary	Durham	Halifax	Kitchener	Markham	RoW	Saskatoon	St.John's	Vancouver	Victoria	Waterloo	Winnipeg
Physical	Joint type								0.10					
	Diameter	0.07	0.07	0.08	0.06	0.06	🟨	0.10	0.06	0.06	0.05	0.07	0.04	0.03
	Material	0.35	0.34	0.32	0.33	0.21	0.40	0.08	0.19	0.11	0.03	0.09	0.14	0.40
	Length	0.22	0.26	0.19	0.34	0.38	0.17	0.35	0.28	0.27	0.43	0.31	0.34	0.30
	Restrained	🟨												
	Roughness									0.11				
	Deadend		0.01											
Historical	Failure month	0.12		0.10	0.10	0.14	🟨	0.24		0.18	0.10	0.18	0.19	0.03
	Install month												🟨	
	Status	0.03		🟨				🟨	0.003					0.003
	Age	0.19	0.32	0.18	0.17	0.22	0.18	0.23	0.37	0.26	0.36	0.29	0.23	0.24
	Replaced status								🟨					
Protection	Casing material	0.02												
	Lining material			🟨	🟨	🟨		🟨	🟨		0.04	0.01	🟨	
	Lining status			🟨	🟨	🟨	🟨	🟨	🟨			🟨	🟨	
	Lining age			0.08		0.003	0.08	🟨	0.005				0.06	
	Cathodic Protection status	0.01		🟨			🟨							
	Cathodic Protection age			0.08			0.17							
	Coating material										🟨			
Operational	Service type	🟨									🟨		🟨	
	Pressure											0.04		
Full model F1		97.5	97.5	97.5	95.3	97.2	98.9	98.6	97.4	95.2	99.4	92.9	97.3	96.7
Final model F1		97.5	97.5	97.5	95.3	97.2	98.9	98.6	97.4	95.2	99.4	92.9	97.2	96.7
Full model Recall		98.4	98.3	98.6	96.7	98.8	99.2	100	98.6	97.6	99.9	96.8	98.5	97.2
Final model Recall		98.6	98.3	98.6	96.4	98.8	99.4	99.9	98.6	97.6	99.9	96.8	98.5	97.2

Due to the significant impacts of the common attributes, i.e., material, diameter, length, and age, according to the results of this analysis as well as their availability in all utilities, these attributes were employed in a separate random forest model to predict the break status. The results are provided in Table 5-9. Comparing the performance of the model with common attributes and the final model indicates the additional attributes in the final model can slightly improve the model's performance. However, the differences are negligible, and using random forest, the common attributes can predict the break status accurately. Hence, according to the results of random forest analysis, utilities that are just beginning to apply predictive models, can develop accurate prediction models by collecting information on common attributes only. In this prediction, material and length are the most contributing factors, and diameter has the lowest contribution in all utilities except the Region of Waterloo. It is noticeable that, unlike other utilities, the Region of Waterloo is the only utility with only 20% of small diameter pipes as presented in data summery table provided in Appendix D.

Table 5-9 Random Forest model results with common attributes

	Common attributes weights				**Common Model**		**Final-Model**	
	Diameter	Material	Length	Age	**F1 Score**	**Recall**	**F1 Score**	**Recall**
Barrie	**0.10**	**0.40**	0.29	0.21	96.8	97.7	97.5	98.6
Calgary	**0.10**	**0.40**	0.29	0.21	97.4	98.3	97.5	98.3
Durham	**0.05**	**0.44**	0.28	0.23	96.3	98	97.5	98.6
Halifax	**0.06**	**0.43**	0.35	0.15	95.2	96.3	95.3	96.4
Kitchener	**0.08**	0.24	**0.43**	0.25	97.1	98.7	97.2	98.8
Markham	**0.03**	**0.55**	0.24	0.17	98	99.2	98.9	99.4
RoW	0.14	**0.13**	**0.47**	0.26	98	99.7	98.6	99.9
Saskatoon	**0.07**	0.26	0.30	**0.37**	97.1	98.5	97.4	98.6
St.John's	**0.08**	0.16	**0.41**	0.35	94.4	96.8	95.2	97.6
Vancouver	**0.08**	0.16	**0.41**	0.35	99.4	100	99.4	99.9
Victoria	**0.08**	0.13	**0.40**	0.38	92.3	96.2	92.9	96.8
Waterloo	**0.08**	0.13	**0.40**	0.38	96.8	97.3	97.2	98.5
Winnipeg	**0.08**	0.13	**0.40**	0.38	96.7	97.2	96.7	97.2

5.3.3.2XGBOOST-RFECV

The overall results of the XGBOOST-RFECV analysis are provided in Table 5-10. Comparing the selected features with the available predictors for each city indicates similar to RF-RFECV, this approach significantly reduces the number of features in the majority of the utilities, and still, the models with the selected features perform equally or slightly better than the full models in terms of F1 and recall scores. Durham and Kitchener are the only utilities with a lower recall score in the selected model than the full model. However, the difference is negligible compared with the high recall score in the full and final model.

According to the results, while in CATPCA analysis, the protection activities' information is the main factor affecting deterioration, this analysis highlights the significant importance of physical and historical predictors. More specifically, the analysis has consistently rated material, diameter, length, age, and failure month as the most important factors. Similar to RF-RFECV, in this analysis material is the most contributing factors among the selected features. In this analysis, information on protection age is consistently important in all utilities except Kitchener, where only less than 1% of the pipes are lined. Also, according to the results of this analysis, the lining material is essential for the utilities with broad lining practices only, i.e., more than 40% of the lined pipes. The analysis has also indicated the significant impacts of some attributes such as restrained, install month, casing material, roughness, and deadend on the pipes' deterioration. As mentioned earlier, these attributes are collected by one utility only, and their actual impacts require further studies.

Table 5-10 XGBOOST-RFECV weights and results - Break Status (Orange - important, yellow – not important, blank – not available)

	Attributes	Barrie	Calgary	Durham	Halifax	Kitchener	Markham	RoW	Saskatoon	St.John's	Vancouver	Victoria	Waterloo	Winnipeg
Physical	Joint type								0.10					
	Diameter	0.06	0.05	0.02	0.02	0.10	0.02	0.20	0.07	0.06	0.10	0.12	0.07	0.02
	Material	0.53	0.66	0.69	0.85	0.59	0.78	0.13	0.52	0.41	0.36	0.34	0.41	0.77
	Length	0.04	0.08	0.04	0.06	0.12	0.04	0.17	0.07	0.09	0.10	0.14	0.12	0.05
	Restrained	0.05												
	Roughness									0.16				
	Deadend		0.01											
Historical	Failure month	0.04		0.03	0.02	0.11	0.02	0.11		0.15	0.12	0.24	0.16	0.01
	Install month												0.02	
	Status	0.07		🟨				0.05	0.04					0.06
	Age	0.05	0.20	0.07	0.03	0.07	0.07	0.11	0.16	0.13	0.12	0.16	0.10	0.09
	Replaced status								🟨					
Protection	Casing material	0.05												
	Lining material			🟨	0.01	🟨		🟨	🟨		0.14	🟨	🟨	
	Lining status			🟨	🟨	🟨	🟨	🟨	🟨			🟨	🟨	
	Lining age			0.12		🟨	0.03	0.23	0.05				0.12	
	Cathodic Protection status	0.09		🟨			🟨							
	Cathodic Protection age			0.02			0.03							
	Coating material										0.05			🟨
Operational	Service type	0.04									🟨		🟨	
	Pressure											🟨		
Full model F1		86.5	91.8	89.7	90.9	79.6	90.3	61.3	88	77	35	69	74	86
Final model F1		86.5	91.8	89.7	90.9	73	90.3	61.3	88	77	27	69	74	86
Full model Recall		84.8	89.3	85.9	89.1	78	86.6	60	84	71.5	27	62	68	83
Final model Recall		84.8	89.3	85.6	89.1	72	86.6	60	84	71.5	24	62	69	84

Similar to previous section, the common attributes were employed in a separate XGBoost model to predict the break status. The results are provided in Table 5-11. As presented in the results, similar to random forest model, the model with selected features slightly outperforms the model with common attributes. As mentioned earlier, utilities at lower level of maturity, who are just beginning to apply predictive models, can develop accurate prediction models information on common attributes only. In this analysis, material and length are the most contributing factors and diameter is the least contributing one in majority of the utilities.

Table 5-11 XGBoost model results with common attributes

	Common attributes weights				**Common Model**		**Final-Model**	
	Diameter	Material	Length	Age	**F1 Score**	**Recall**	**F1 Score**	**Recall**
Barrie	0.08	**0.78**	**0.06**	0.08	85.9	84.8	86.5	84.8
Calgary	0.079	**0.78**	**0.06**	0.08	85.9	84.8	91.8	89.3
Durham	**0.033**	**0.83**	0.049	0.09	88	84	89.7	85.6
Halifax	**0.030**	**0.86**	0.064	0.05	91	89	90.9	89.1
Kitchener	0.12	**0.65**	0.14	**0.09**	78	71	73	72
Markham	**0.032**	**0.82**	0.06	0.09	90	87.5	90.3	86.6
RoW	**0.08**	0.16	**0.66**	0.09	58.8	52.6	61.3	60
Saskatoon	**0.08**	**0.64**	0.088	0.19	86.2	81.8	88	84
St.John's	**0.09**	**0.65**	0.1	0.16	72.5	66	77	71.5
Vancouver	**0.14**	**0.39**	0.21	0.26	24	17.5	27	24
Victoria	**0.14**	**0.45**	0.19	0.22	68.5	61.6	69	62
Waterloo	**0.11**	**0.62**	0.15	0.13	66.5	62.6	74	69
Winnipeg	**0.016**	**0.86**	0.047	0.08	85.6	83.5	86	84

Comparing results of this analysis with random forest indicates in general, random forest models are performed better than XGBOOST models for this target.

6 Discussion

The results of the study will be discussed in three levels: applicability of the applicability of the approaches, the important features, and limitation and future recommendation. Each level is explained in detail in the following paragraphs:

6.1 Applicability of the dimensionality reduction approaches

Although the mentioned approaches follow the same goal of dimensionality reduction, they are different in nature. When data are all numerical and linearly related to the target, correlation coefficients and the approaches based on correlation are good indicators representing the most significant factors influencing the target. Besides correlation analysis, Factor Analysis of Mixed Data (FAMD) that applies to mixed types of data is recommended when there are linear relationships between target variables and predictors and categorical variables with an equal level of categories. In contrast, Categorical Principal Component Analysis (CATPCA), also known as non-linear PCA, can handle a linear and non-linear relationship among variables and is recommended when mixed types of data in the analysis are not linearly related to the target. Recursive Feature Elimination Approach with Cross-Validation (RFECV) is another feature selection approach that can handle different relations between the predictors and targets through a proper estimator selection. In this analysis, selecting an estimator that best fits the data and tunning its hyperparameters determines the reliability of the results. In practice, the number of available estimators is significant and choosing an appropriate estimator based on data structure, and the predictors' relations with the target is critical. For instance, the simple linear and regularized linear regressors such as elastic net regressors were initially selected as the estimator in this study. However, the resulting negative R-squared indicated the inability of these estimators to fit in the data and select the most important features accordingly. Hence, later on, XGBOOST and Random Forest were tested out. With resulting high performance, the two algorithms were employed instead. Compared with XGBOOST, the Random Forest estimator was fitted better in the data and resulted in more accurate models. Accordingly, although the selected features by the two approaches were almost similar, the results of RF-RFECV seem to be more reliable.

The identified important factors and the most contributing factors vary between different models based on the mentioned differences in the current study. More specifically, while physical and historical information were the most contributing factors in RFECV approach, in CATPCA analysis the protection activities were contributing the most. Comparing different approaches of this study indicates, while CATPCA is a flexible approach in which implementation is straightforward, in RFECV, selecting a proper estimator and tuning its hyperparameters is critical and time-consuming. Also, since the categorical variables have a different number of category levels herein, and due to the lack of linear relationship between the predictors and targets, the results of FAMD analysis might be biased. Comparing the approaches based on the mentioned differences and similarities, it seems that CATPCA is the most reliable approach for the current study.

6.2 Important attributes

The findings of this study highlight the impacts of physical characteristics such as diameter, material, and length on pipes' current rate of failure and break status. All utilities consistently collected these variables. As mentioned earlier in the literature, these factors were commonly used in previous watermain prediction models. Hence, their level of significance in the results of this study is aligned with previous findings and was expected.

The results also rated dead-end as another important physical factor affecting deterioration. Dead-end pipes are referred to the segments that are closed on one side. Jim Angres (2002), in a study on best practices for pipes installation, warned about the dead-end installation of the pipes. According to the study, decreasing flow results in stagnation in dead-end lines, affecting water quality and leading to corrosion. To the best of our knowledge, this attribute was not available in the previous watermain deterioration studies. However, according to this study's results, this attribute's importance confirms the suggested best pipe installation practices by Jim Angres (2002).

Also, as mentioned in the literature, joint failure is a common failure type in water networks (Burn, et al., 2005) (Dingus, et al., 2002) which are promoted by the pipe joint type. For instance, since rigid joints cannot handle ground movement, they are highly prone to leak and fracture failures. In this study, different types of rigid, e.g., mechanical joints and flexible joints, e.g., rubber, were available in the analysis. Hence, although this approach has not been employed in previous pipe deterioration models, identifying joint type as an influential factor in this analysis was expected.

Restrained pipes refer to the additional structural supports on the pipes controlling movements and preventing axial displacement or flexure at bends. Hence, restraining the pipes exposed to water hammers or thrust forces will strengthen pipes and prevent failures. Restraining is considered a physical factor in watermain data, and pipes can be restrained through mechanical joints, adding a concrete block, also known as thrust blocks, at the end of line or beside a joint, and grip pipes using coupling restraints. To the best of our knowledge, previous studies have not employed this factor as a predictor of watermain deterioration models. This attribute is available for one utility, and the results of this study also indicate its importance in the XGBOOST-RFECV model only. It is important to focus on the results of a more reliable approach, i.e., CATPCA. Hence, the actual impact of this attribute requires further investigation.

Roughness as another physical factor measures the irregularities in the inner surface of the pipes. Pipe inner surface irregularities depend on the pipe material and can change over time. This attribute is also recorded by a few cities only, and its impacts on deterioration were not revealed in previous studies. Hence, although the current study highlights the impacts of roughness on pipe failures, further research is required to support this.

The literature discussed the impact of pressure on the pipe, and the importance of this attribute in the analysis can be explained accordingly. However, this attribute was recorded by a few utilities only, and generalization of the impacts requires further data and supports from future research.

As the only environmental factor, the soil type was available for broken pipes only. This analysis highlights the impacts of this attribute on the pipe failure rate. As mentioned in the literature, soil type can reflect different soil characteristics such as PH and humidity and can affect pipe failure directly, i.e., soil movement, or indirectly, i.e., corrosion. For the current study, the soil type classified soils as either one type, i.e., clay, sand, gravel, etc., or a mix of different soil types. Although this study indicates the importance of soil type in pipes' current failure rate, it cannot answer how failure rates change in different soil types.

As a historical factor, the age of the pipes was commonly employed in previous pipe prediction models to describe the time dependency of breakage and estimate an optimal time for pipe replacement. Also, based on the results of the previous studies, once a pipe is failed for the first time, it continues to be failed in the other location. Changes in rate and probability of failure throughout the age and significant dependency of pipe failure status on the previous failure were also observed in this study's results which support the literature. Also, as mentioned earlier in the literature, pipes usually follow a typical life cycle represented by a “bathtub curve”. The bathtub curve describes the period right after installation with failures due to defective pipes or installation problems, a trouble-free period, and a period of increasing breakage rate due to aging and deterioration around the pipe material expected service life. The early failures were also observed in the results of this study, and to explain them according to the bathtub curve, records of day and month of installation and failure are required. Besides, failure month itself was identified as a key factor affecting pipe deterioration in this analysis. In this study, installation month, extracting from installation date, indicates at which month of the year the pipe was installed.

This study also pointed out significant impacts of protection activities on pipe deterioration. According to the results, not only the protection status is important, and the protected pipes are less prone to failure, but also other characteristics of protection, including the material used for that and its age, are important factors affecting failures. These results confirm the literature about the impacts of lining on increasing pipes' expected service life. Surprisingly, the results of correlation analysis and consistently identifying lining age for all utilities compared with age in different analyses. While previous studies have mainly focused on pipe age for predicting watermain failure, lining and protection age are more influential factors. Although the study highlights the significant impacts of protection activities on deterioration, it is unclear how different protection materials and age of protection can affect failures.

6.3 Limitations, recommendations, and potential future research

A wide range of input variables was available in this study. Some of the collected variables consist of more than 20% missing values. Replacing a large percentage of missing values based on the mentioned approaches in methodology can lead to unrealistic results. Hence, those attributes were removed from the analysis, and their impacts of deterioration remained ambiguous. The provided attributes were varied by the utilities. Thus, although the same dimensionality reduction approaches were applied to each city, the results were not easily comparable.

Besides, some recorded information was not consistently collected for both broken and non-broken pipes. It limits the applicability of those attributes on break status analysis, where records of both broken and non-broken pipes are required. In particular, information on soil type was recorded for broken pipes only, and the impacts of this attribute on break status analysis remained unspecified.

One of the main challenges in this study was related to the replaced pipes. In some utilities, records of the broken pipes that were replaced during the maintenance were removed from the inventory, and the new pipe was assigned the same pipe ID as the replaced broken pipe. Accordingly, it led to negative ages for broken pipes and, in some cases, miss-matches of material and/ or diameter in break and inventory files. The miss-matches and negative ages were discussed with the cities and addressed accordingly.

Another challenge in the analysis was related to the small number of broken pipes, resulting in the small training size for the RFECV approach and the unbalanced dataset in break status analysis. Training models with small training sizes is challenging and might lead to under-fitted models and poor approximation. However, this is related to the nature of this study and cannot be avoided.

For the current study, more than 70% of the time was on the data cleaning process. This process was related to identifying gaps and inconsistencies in data and addressing them. Inconsistency in data mainly was related to different formats of collecting categorical variables and different units of measurement for numerical predictors. For instance, the material was recorded by its full name, abbreviation, or code in and within utilities, and diameter was recorded in a different unit of measurement, i.e., mm, in, etc.

Also, as mentioned earlier, information on adjacent assets that can be extracted from GIS is the most reliable approach for replacing the missing values. However, GIS files were available only for a few utilities, and the approach could not be applied broadly to replace the missing values.

Therefore, to address the mentioned limitation, improve the current practice, and shorten future data cleaning processes, the following points can be taken into consideration:

- When recording the variables, a unique name and appropriate level of measurement should be defined for categorical and numerical variables.

- To avoid miss-matches and negative ages, the replaced pipes should assign a new pipe ID, and the record of the broken pipes should be preserved in the data.

- Consistently recording data enhances future studies to evaluate the impacts of different factors in broken and non-broken pies. In particular, as mentioned earlier, soil type, which was also identified as an important factor in the current rate of failure analysis, was not available for break status.

- As mentioned earlier, GIS is a reliable source for replacing missing data. It is recommended that utilities record all the data and update the GIS file regularly. So that future studies can compute the missing information more accurately.

- Future studies can also focus on identifying impacts of protection activities on durability and costs and identifying effects of operational activities on durability

- Since soil type was an important factor in this study, it is worth future studies to integrate and incorporate other soil characteristics such as soil PH, resistivity, corrosivity, etc., and other environmental factors, including weather, in records of data and evaluate their impacts on watermain deterioration.

- Lastly, the pipe wall thickness and manufacturing defects are other mentioned factors in the literature that the current study did not have reliable data about them. Future studies should also leverage them in watermain prediction models and evaluate their impacts accordingly.

7 Summary and Conclusion

Aging water infrastructure is a worldwide concern that can jeopardize water systems' ability to deliver clean water safely. Better understanding the factors that lead to water infrastructure failure is key to better managing these critical assets. The present thesis focused on applying different dimensionality reduction approaches to evaluate the impacts of a broad range of factors on deterioration and identify the most important factors affecting deterioration. Data was collected from thirteen Canadian utilities, including Barrie, Calgary, Durham, Halifax, Kitchener, Markham, Region of Waterloo, Saskatoon, St. John's, Vancouver, Victoria, Waterloo, and Winnipeg.

As discussed earlier, out of the selected dimensionality reduction approaches, CATPCA is the most reliable and straight forward approach. Accordingly, a data collection framework can be organized into three different levels based on the results of CATPCA and a few findings from RFECV approaches. The first level represents the minimum level of the required data for reliable prediction models. These data include the most common physical and historical data such as material, diameter, length, age, and number of failures. In the second level, utilities that already have the minimum required data can focus on collecting data on protection activities on pipes. More specifically, data on type and date of protection. Lastly, the third level could be one step further on collecting the factors that were available for a few utilities only. The factors include environmental, operational, and some of the physical factors such as joint type, roughness, dead-end, restrained, and pipe depth. The final important factors at each level are listed in Table 7-1. The attributes at each level are highlighted in yellow, green, and red respectively.

The data in the first level were available for all utilities and identified as the topmost important factors in utilities without data on protection activities based on the results of CATPCA. Also, as demonstrated in Table 5-9. The second level data is related to information on type and date of protection. Where available, these factors were consistently identified as the most important factors affecting the failures. A few of the data in the third level, i.e., roughness and restrained, were identified as important by RFECV approaches for break status target, not the CATPCA. However, the others were suffering from either availability for one target, i.e., soil type, or were not consistently identified as important by CATPCA. Hence, as one step further, it is worth it for utilities to start recording information on them so that future research can evaluate their actual impacts on failures.

In general, utilities that are just beginning to apply predictive models can develop accurate prediction models by collecting information on common attributes only. However, the second level can help more advanced utilities improve their current predictions and maintenance plans. For the third level data, the results of this study could not reveal their impacts on degradation models. Hence, it is worth it for more advanced utilities to collect that information and evaluate their impacts on their prediction models and maintenance practices.

Also, as a part of the suggested data collection framework, these data should be collected in a correct structure to avoid the long future data cleaning process. More specifically, a consistent format of recording categorical and date variables and a unique unit of measurement for the numerical should be defined in advance. The specified format after the data cleaning process for the available variables in this study is provided in Appendix E.

Table 7-1 List of important factors (First level is highlighted in yellow, second level in green, and third level in red)

Physical	**Historical**	**Protection activities**	**Operational**	**Environmental**
Material	Full installation date	Cathodic protection year	Pressure	Soil type
Diameter	Full failure date	Lining Material	Service Type	-
Length	Status	Lining Year	-	-
Join type	-	Coating Material	-	-
Roughness	-	Anode type	-	-
Dead-end	-	-	-	-
Restrained	-	-	-	-
Pipe Depth	-	-	-	-

References

Carreira-Perpinan, M. (1997). *A Review of Dimension Reduction Techniques.*

U. S. Environmental Protection Agency. (2002). *Deteriorating Buried Infrastructure. Management Challenges and Strategies.* United State.

Ambielli, B. (2017, Octoober 29). *Gini Impurity (With Examples)*. Retrieved from Bambielli's Blog: https://bambielli.com/til/2017-10-29-gini-impurity/#:~:text=Gini%20Impurity%20is%20a%20measurement,labels%20from%20the%20data%20set.

Andreou, S. M. (1987b). A new methodology for modelling break failure patterns in deteriorating water distribution systems: applications. *advance in water resources, 10*(1), 11-20. doi:10.1016/0309-1708(87)90003-0

Andreou, S.A., Marks, D.H., Clark, R.M. (1987a). A new methodology for modelling break failure patterns in deteriorating water distribution systems: theory. *Advance in water, 10*(1), 2-10. doi:10.1016/0309-1708(87)90002-9

Aslani, B., Mohebbi, S., & Axthelm, H. (2021). Predictive analytics for water main breaks using spatiotemporal data. *Urban Water, 18*(6), 433-448. doi:https://doi.org/10.1080/1573062X.2021.1893363

Asnaashari, A., McBean, E. A., Shahrour, I., & Gharabaghi, B. (2009). Prediction of watermain failure frequencies using multiple and Poisson regression. *Water Supply, 9*(1), 9-19. doi:10.2166/ws.2009.020

Bardakjian, H., McReynolds, M., & Hausmann, D. (2007). Corrosion Protection of Large Diameter Welded Steel Pipelines with Cement Mortar Coatings. *International Conference on Pipeline Engineering and Construction*, (pp. 1-10). Boston. doi:10.1061/40934(252)94

Barton, A. N., Farewell, S. T., Hallett, H. S., & Acland, F. T. (2019). Improving pipe failure predictions: Factors affecting pipe failure in drinking water networks. *Water Research, 164*, 114926–114926. doi:https://doi.org/10.1016/j.watres.2019.114926

Berardi, L., Giustolisi, O., Kapelan, Z., & Savic, D. (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

Brito, P. (2014). Symbolic Data Analysis: another look at the interaction of data mining and statistics. *Wiley interdisciplinary reviews-data mining and knowledge discovery, 4*(4), 281-295. doi:10.1002/widm.1133

Brown, L. (2006). *Winter considerations: Ice formation, freezing index, and frost penetration.* British Columbia: Ministry of Agriculture and Lands.

Bruaset, S., & Sægrov, S. (2018). An analysis of the potential impact of climate change on the structural reliability of drinking water pipes in cold climate regions. *Water, 10*(4). doi:https://doi.org/10.3390/w10040411

Burn, S., Davis, P., Schiller, T., Tiganis, B., Tjandraatmadja, G., Cardy, M., . . . Whittle, A. (2005). *Long-term Performance Prediction for PVC Pipes.* AWWRF.

Cadima, J., & Jolliffe, I. (2009). On relationships between uncentred and column-centred principal component analysis. *Pakistan journal of statistics, 25*(4), 473-503.

Canadian Infrastructure Report Card (CIRC). (2012). *2012 Municipal Roads and Water Systems.*

Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S., & Zhou, S. (2019). A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. *Diagnostics, 9*(4). doi:https://doi.org/10.3390/diagnostics9040178

Chen, M.-H., Shao, Q.-M., & Ibrahim, J. (2000). Bayesian Variable Selection. In Springer (Ed.), *Monte Carlo Methods in Bayesian Computation* (pp. 267-306). New York, NY: Springer Series in Statistics. doi:https://doi.org/10.1007/978-1-4612-1276-8_9

Chen, T., & Guestrin, C. (2016). doi:10.1145/2939672.2939785

City of Toronto. (2020). *Watermains*. Retrieved from Toronto: https://www.toronto.ca/services-payments/building-construction/infrastructure-city-construction/understanding-city-construction/water-sewer-mains/

Cooper, L., Schwartz, D., & Reamon, D. (2012). *Using Random Forests to Identify Factors of Student Motivation in a Project-Based Learning Course* (Vol. 5). doi:10.1115/IMECE2012-86088

Date, S. (2019, November 9). *Toward data science*. Retrieved from The Akaike Information Criterion: https://towardsdatascience.com/the-akaike-information-criterion-c20c8fd832f2

Dingus, M., Haven, J., Austin, R., & AWWA Research Foundation. (2002). *Nondestructive, noninvasive assessment of underground pipelines.* AWWA Research Foundation and American Water Works Association.

Fang, G., Liu, W., & Wang, L. (2020). A machine learning approach to select features important to stroke prognosis. *Computational Biology and Chemistry, 88*. doi:https://doi.org/10.1016/j.compbiolchem.2020.107316

Folkman, S. (2018). *Water main break rates In the USA and Canada: a comprehensive study.* Utah State University.

Frost, J. (2020). *Guide to Stepwise Regression and Best Subsets Regression*. Retrieved from Statistics by Jim: https://statisticsbyjim.com/regression/guide-stepwise-best-subsets-regression

Gioia, F., & Lauro, C. N. (2006). Principal component analysis on interval data. *Computational statistics*, 343–363.

Giraldo, M., & Rodríguez Sánchez, J. (2020). Comparison of Statistical and Machine Learning Models for Pipe Failure Modeling in Water. *Water, 12*(4), 1153–1153. doi:https://doi.org/10.3390/w12041153

Granitto, P. M., Cesare, F., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems, 83*(2), 83-90. doi:https://doi.org/10.1016/j.chemolab.2006.01.007

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430-450. doi:10.1037/1082-989X.6.4.430

Guan, S. W. (2001). Corrosion protection by coatings for water and wastewater pipelines. *Appalachian Underground Corrosion Short Course, Water and Wastewater Program*. West Virginia University, PA.

Guyon, I., Weston, J., Barnhill, S., & Vapnik , V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *46*, 389–422. doi:10.1023/A:1012487302797

Hayes, A. (2020, March 18). *R-Squared Definition*. Retrieved from Investopedia: https://www.investopedia.com/terms/r/r-squared.asp

Hong, H. (1998). Reliability based optimal inspection schedule for corroded pipeline. *Annual conference of the canadian society for civil engineering*, (pp. 743-752). Halifax, NS, Canada.

Hu , Y., & Hubble, D. W. (2005). Failure conditions of asbestos cement water mains in Regina. *Canadian Society of Civil Engineering (CSCE)* (pp. 1-10). Toronto: National Research Council Canada.

Hu, Y., & Hubble, W. D. (2007). Factors contributing to failure of asbestos cement water mains. *Canadian journal of civil engineering, 34*(5), 608–621. doi:https://doi.org/10.1139/L06-162

Hudson, W., & Haas, R. (2013). *Public Infrastructure Asset Management, second edition.* McGraw-Hill Publishing.

Jaadi, Z. (2019, September 4). *data science step step explanation principal component analysis.* Retrieved from builtin: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Jolliffe, I. (2002). *Principal component analysis.* New York: Springer-Verlag.

Jolliffe, I., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactionns of the roya society a-mathematical physical and engineering science, 374*(2065). doi:10.1098/rsta.2015.0202

Jun, H., Park, J., & Bae, C. (2020). Factors Affecting Steel Water-Transmission Pipe Failure and Pipe-Failure Mechanisms. *Environmental Engineering, 146*(6). doi:https://doi.org/10.1061/(ASCE)EE.1943-7870.0001692

Karimian, S. F. (2016). *Failure rate prediction models of water distribution networks.* Montreal: Concordia University.

Karsten, B. (2010). *Stochastic differential equations: An introduction with applications.* Springer science & business media.

Kettler, A. J., & Goulter, I. C. (1985). An analysis of pipe breakage in urban water distribution networks. *Canadian journal of civil engineering, 12*(2), 286-293. doi:10.1139/l85-030

Kirmeyer, G. J. (1994). *An Assessment of Water Distribution Systems and Associated Research Need.* Foundation and American Water Works Association.

Kleiner, Y., & Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: Statistical models. *Urban Water, 3*(3), 131-150.

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models.* CRC Press.

Kuraoka, S., Rajani, B., and Zhan, C;. (1996). Pipe-soil interaction analysis of field tests of buried PVC pipe. *Infrastructure systems, 2*(4), 162 - 170. doi:10.1061/(ASCE)1076-0342

Li, L. (2015). *Selected Applications of Convex Optimization.* New York: Springer.

Linting, M., & van der Kooij, A. (2012). Nonlinear principal components analysis with CATPCA: a tutorial. *Journal of Personality Assessment, 94*(1), 12-25. doi:10.1080/00223891.2011.627965

Linting, M., Meulman, J. J., Groenen, P. J., & Koojj, A. J. (2007). Nonlinear principal components analysis: Introduction and application. *Psychological Methods, 3*(12), 336–358. doi:https://doi.org/10.1037/1082-989X.12.3.336

Malizio, A. B. (1986). Pipe digs show effectiveness of poly sheet encasement. *Water/Engineering & Management*. doi:133: 3: 27:28

Mallawaarachchi, V. (2017, July 7). *towards data scienc*. Retrieved from introduction to genetic algorithms: https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3

Manitoba Soils and Screw Piles. (2018, September 28). *Winnipeg's Frost Line Can Be As Deep As 8 Feet*. Retrieved from manitoba screw piles: https://www.manitobascrewpiles.ca/winnipegs-frost-line-8-feet/

Mark, A. J. (1985). Statistical models for water main failures. *US Environmental Protection Agency (Co-operative Agreement CR810558) M.I.T Office of Sponsored Projects.* Boston.

Mitsuhiro, M., & Yadohisa, H. (2018). A unified representation of simultaneous analysis methods of reduction and clustering. *Japanese Journal of Statistics and Data Science, 1*(2), 393-412. doi:https://doi.org/10.1007/s42081-018-0022-6

Municipal Association of South Carolina (MASC). (2016, March). *Life expectancy of water lines*. Retrieved from Municipal Association of South Carolina: https://www.masc.sc/Pages/newsroom/uptown/March-2016/Life_expectancy_water_lines.aspx

Nain, S., Garg, D., & Kumar, S. (2018). Performance evaluation of the WEDM process of aeronautics super alloy. *Materials and Manufacturing Processes, 33*, 1-16. doi:10.1080/10426914.2018.1476761

Najjar, Y., & Basheer, Y. (1996). Neural network approach for site characterization and uncertainty prediction. *ASCE Geotechnical Special Publication*, (pp. 134–148).

Nanan, K. (2019, January 7). *Your Comprehensive Guide to Waterproofing with Bituminous Paint*. Retrieved from corrosionpedia: https://www.corrosionpedia.com/your-comprehensive-guide-to-waterproofing-with-bituminous-paint/2/6928

Nathanson, J. A. (2020, March 31). Water supply system. *Encyclopædia Britannica*. Encyclopædia Britannica, inc. Retrieved from Encyclopædia Britannica: https://www.britannica.com/technology/water-supply-system

Nishiyama, M., & Filion, Y. (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

Pagès, J. (2014). *Multiple Factor Analysis by Example Using R.* (C. a. Hall/CRC, Trans.)

Rahman, S., & Farrell, H. (2007). *Municipal Pipeline Thrust Restraint: Next Generation of Products for Thermoplastic Pipes.* Fort Worth, Texas, USA.

Rajani, B., & Kleiner, Y. (2003). Protecting ductile iron water mains: what protection method works best for what soil condition? *American Water Works Association, 95*(11), 110-125. doi:10.1002/j.1551-8833.2003.tb10497.x

Rajani, B., & Zhan, C. (1996). On the estimation of frost loads. *Canadian geotechnical journal., 33*(4), 629-641. doi:10.1139/t96-088-309

Rajeev, P., Kodikara, J., Robert, D., & Zeman, P. (2015, january). Factors contributing to large diameter water pipe failure. *Water Asset Management International, 10*, 9-14.

Ramsay, J., & Silverman, B. W. (2005). *Functional data analysis.* New York: Springer-Verlag.

Roughgarden, T., & Valiant, G. (2015, April 27). *Lecture #9: The Singular Value Decomposition (SVD) and Low-Rank Matrix Approximations.* Retrieved from CS168: The Modern Algorithmic Toolbox: http://theory.stanford.edu/~tim/s15/l/l9.pdf

Saskatchewan Water Security Agency. (2004). *Water Pipeline Design Guidelines.* Saskatchewan.

Sharma, A. (2020, June 30). *4 Simple Ways to Split a Decision Tree in Machine Learning*. Retrieved from analytics vidhya: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/

Shlens, J. (2014). A tutorial on principal component analysis. *International Journal of Remote Sensing*.

Snider, B., & McBean, E. (2020). Improving Urban Water Security through Pipe-Break. Prediction Models: Machine Learning or Survival Analysis. *Environmental Engineering, 146*(3). doi:10.1061/(ASCE)EE.1943-7870.0001657

Snider, B., & McBean, E. (2020, October). State of Watermain Infrastructure: A Canadian Case Study using Historic Pipe Break Datasets. *Canadian Journal of Civil Engineering*. doi:10.1139/cjce-2020-0334

Snider, B., & McBean, E. (2020). Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions. *Urban Water Journal, 17*(2), 163-176. doi:https://doi.org/10.1080/1573062X.2020.1748664

Stephanie, G. (2020, March 5). *Mallows' Cp*. Retrieved from Elementary Statistics for the rest of us!: https://www.statisticshowto.com/mallows-cp/

Stepwise regression. (n.d.). Retrieved from NCSS Statistical Software: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise_Regression.pdf

Town of Cobalt. (2014). *Asset management plan.*

U. S. Environmental Protection Agency. (2010). *Control and Mitigation of Drinking Water Losses in Distribution.* Washington: Bibliogov.

U. S. Environmental Protection Agency. (2012). *Condition assessment technologies for water transmission and distribution systems.* United States: Bibliogov.

US Army Corps of Engineers. (1998, March 31). *EngineerManuals.* Retrieved from US Army Corps of Engineers: https://www.publications.usace.army.mil/Portals/76/Publications/EngineerManuals/EM_1110-2-2902.pdf

Vakhania, N., Tarieladze, V., & Chobanyan, S. (1978). Covariance Operators. In: Probability Distributions on Banach Spaces. *Theory of Probability & Its Applications, 14*, 144-183. doi:https://doi.org/10.1007/978-94-009-3873-1_3

Vasudevan, D. S. (2014). How does one calculate factor score in factor analysis? Retrieved from https://www.researchgate.net/post/How_does_one_calculate_factor_score_in_factor_analysis/53c50b3fd5a3f2140c8b465e/citation/download.

Vidal, R., Ma, Y., & Sastry, S. (2016). *Generalized principal component analysis.* New York: Springer.

Voyle, N., Keohane, A., Newhouse, S., Lunnon, K., Johnston, C., Soininen, H., . . . Dobson, R. J. (2016). A Pathway Based Classification Method for Analyzing Gene Expression for Alzheimer's Disease Diagnosis. *Journal of Alzheimer's disease, 49*(3), 659–669. doi:10.3233/JAD-150440

Walski , T. M., & Pelliccia, A. (1982). Economic analysis of water main breaks. *American Water Works Association, 74*(3), 140-147. doi: https://doi.org/10.1002/j.1551-8833.1982.tb04874.x

Wang, S., & Chen, S. (2019). Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. *Journal of Petroleum Science and Engineering, 174*, 682-695. doi:https://doi.org/10.1016/j.petrol.2018.11.076

Wang, Y., Zayed, T., & Moselhi, O. (2009). Prediction Models for Annual Break Rates of Water Mains. *Journal of Performance of Constructed Facilities, 23*(1), 40-46. doi:https://doi.org/10.1061/(ASCE)0887-3828(2009)23:1(47)

Why should we avoid dead ends? (2002). *Opflow, 28*(10), 10–11. doi:https://doi.org/10.1002/j.1551-8701.2002.tb01678.x

Wiley, D. (2018). *Will Lining my Plumbing with Epoxy Coating Resist Future Corrosion?* Retrieved from NuFlow Midwest: https://www.nuflowmidwest.com/will-lining-my-plumbing-with-epoxy-coating-resist-future-corrosion/

Wols, B. A., & Thienen, P. v. (2014). Impact of weather conditions on pipe failure: a statistical analysis. *Journal of Water Supply, 63*(3), 212–223. doi:10.2166/aqua.2013.088

Yamijala, S., Guikema, S. D., & Brumbelow, K. (2009). Statistical models for the analysis of water distribution system pipe break data. *In reliability engineering and system safety, 94*(2), 282-293. doi:10.1016/j.ress.2008.03.011

Yuichi , M., Masahiro , K., & Naomichi , M. (2016). *Nonlinear Principal Component Analysis and Its Applications.* Singapore: Springer.

Zaiontz, C. (2020). *Everything you need to perform real statistical analysis using Excel*. Retrieved from Real Statistics: http://www.real-statistics.com/multivariate-statistics/factor-analysis/principal-axis-method/

Zeileis, K. J. (2008). Regression models for count data in R. *Journal of statistical software*.

Zhan, C., & Rajani, B. (1997). Estimation of frost load in a trench: theory and experiment. *Canadian geotechnical journal, 34*(4), 568-579. doi:10.1139/t97-023

Appendix A

A 1 Required parameters in RFECV approach

Estimator: A supervised algorithm with a fit method employed to identify important features. This study selected two estimators for each target attribute to determine the most important features, random forest and extreme gradient boosting. These estimators were tuned prior to RFECV.

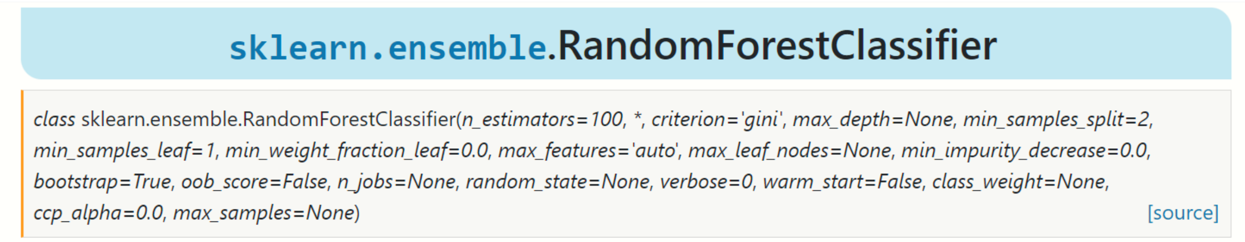
Step: specifies the number of features to remove in each step. For this study, the default value of 1 was kept.

Min_feature_to_select: Specifies the minimum number of features selected by the estimator. The default value of 1 was kept for this study.

Cv: An integer value that specifies the strategy of cross-validation strategy. The default value of None equivalent to 5-fold cross-validation was employed for this study.

A 2 Random forest hyperparameters (Classification and regression)

The complete list of random forest hyperparameters is provided in Figure A - 1 for categorical target variables. The list is the same for random forest regression as well. While the main parameters are optimized to avoid overfitting, the default values are used for the remaining. The optimized parameters are listed below:



sklearn.ensemble.RandomForestClassifier

```
class sklearn.ensemble.RandomForestClassifier(n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2,
min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0,
bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None,
ccp_alpha=0.0, max_samples=None)
```

[source]

Figure A - 1 Hyperparameters of random forest classifier adapted from (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

N_estimators (default=100): The number of estimators indicates the number of employed trees for learning the base model. Since the most important features are selected depending on the majority of the tree's votes, finding an optimal number of estimators is essential. An array of numbers between 10 to 200 for the numerical target attribute and 100 to 200 for categorical was defined for this parameter.

Max_depth (default=none): The max depth specifies how deep or to what level each tree of the forest should be split. The value of this parameter significantly affects the fit of the model. The default value of max_depth split the trees up to a depth where all nodes achieved maximum purity or until samples in each node was fewer than what was defined for min_sample_split. Also, a large value of the depth results in overfitting of the model. For this study, an array of integer numbers between 4 to 22 was defined to find the optimum number.

Max_features (default=”auto”): This parameter specifies the number of variables for the splitting process. The possible values for this parameter are auto, sqrt, and log2. Auto is equivalent sqrt, and when max_fetures is set as sqrt, the number of variables is the square root of the total number of variables during the splitting process. For this study, a list of the three mentioned values is specified for finding the optimal value through RandomizedSearchCV.

Creation: Creation measures the quality of the splits and how pure the splits are. For break status, the default value is ‘Gini,’ and for regression, it is ‘squared_error.’ While for the regression, the default value was kept, for the categorical target variables, two possible values of ‘Gini’ and ‘entropy’ were specified for using in RandomizedSearchCV.

Bootstrap (default=True): Bootstrap specifies if a full data set should be used for building each tree or a bootstrap sample of the dataset. Due to a large amount of data for the break status target attribute, the default value was kept for the analysis. A two-dimension array of True and False was specified for the numerical target to avoid overfitting.

Table A - 1 Random Forest classifier hyperparameters (Full model)

Utility	Random forest classifier hyperparameters (full model)			
	N_estimators	Max_depth	Max_features	Creation
Barrie	180	19	Log2	entropy
Calgary	140	17	auto	entropy
Durham	105	22	auto	gini
Halifax	150	13	auto	gini
Kitchener	135	11	Log2	gini
Markham	130	17	sqrt	gini
Region of Waterloo	155	16	auto	gini
Saskatoon	100	16	auto	entropy
St.John's	155	16	auto	gini
Vancouver	150	13	auto	gini
Victoria	180	18	sqrt	entropy
Waterloo	135	11	Log2	gini
Winnipeg	180	19	Log2	entropy

Table A - 2 Random Forest classifier hyperparameters (model with selected features)

Utility	Random forest classifier hyperparameters (model with selected features)			
	N_estimators	Max_depth	Max_features	Creation
Barrie	100	18	Log2	gini
Calgary	100	16	auto	entropy
Durham	160	18	sqrt	entropy
Halifax	150	13	auto	gini
Kitchener	135	11	Log2	gini
Markham	130	17	sqrt	gini
Region of Waterloo	195	18	sqrt	entropy
Saskatoon	115	17	auto	entropy
St.John's	145	11	auto	entropy
Vancouver	140	17	auto	entropy
Victoria	130	19	auto	entropy
Waterloo	145	11	auto	entropy
Winnipeg	180	19	Log2	entropy

Table A - 3 Random Forest regressor hyperparameters (full model)

Utility	Random forest regressor hyperparameters (full model)			
	N_estimators	Max_depth	Max_features	Bootstrap
Barrie	30	16	auto	False
Calgary	60	15	auto	True
Durham	200	9	auto	True
Halifax	180	9	auto	False
Kitchener	70	14	log2	True
Markham	40	10	auto	True
Region of Waterloo	180	9	auto	False
Saskatoon	50	6	auto	False
St.John's	80	5	auto	True
Vancouver	50	19	auto	True
Victoria	200	9	auto	True
Waterloo	60	15	auto	True
Winnipeg	60	15	auto	True

Table A - 4 Random Forest regressor hyperparameters (model with selected features)

Utility	Random forest regressor hyperparameters (model with selected features)			
	N_estimators	Max_depth	Max_features	Bootstrap
Barrie	180	180	180	180
Calgary	60	60	60	60
Durham	200	200	200	200
Halifax	128	128	128	128
Kitchener	60	60	60	60
Markham	80	80	80	80
Region of Waterloo	80	80	80	80
Saskatoon	50	50	50	50
St.John's	50	50	50	50
Vancouver	140	140	140	140
Victoria	200	200	200	200
Waterloo	200	200	200	200
Winnipeg	200	200	200	200

A 3 Extreme gradient boosting hyperparameters (Classification and regression)

The most important hyperparameter of the XGBOOST method are listed below:

Min_child_weight (default=1): This parameter specifies the minimum weights of the required samples in a child. Specifying this value avoids further partitioning when the sum of the instances' weights in a leaf is reduced to less than min_child_weight. The range of this parameter is between 0 to infinity. For the current study, an array of numbers between 0.001 to 0.1 was defined.

Booster (default=gbtree): This parameter determines the type of learner for the partitioning process. It can be a tree-based function or a linear one. In tree-based, the model consists of groups of trees, while in linear booster, it is a weighted sum of linear functions. The default parameter was kept for this study.

Eta (default=0.3): Eta is a learning rate and indicates the shrinkage of each step one makes. This value can range between 0 to 1. For instance, 1 step at a learning rate of 0.25 makes the weight of the step 0.25. For this study an array of [0.001,0.01,0.02,0.1,0.25,0.5,1] shape was selected for tunning by RandomizedSearchCV.

Lambda (default=1): Lambda is L1 linear parameter on weights ranging between 0 to 1. For this study, an array of [0.001,0.01,0.02,0.1] shape was selected for tunning by RandomizedSearchCV.

Alpha (default=0): Alpha is the L2 linear parameter on weights ranging between 0 to 1. Similar to lambda for this study, an array of [0.001,0.01,0.02,0.1] shape was selected for tunning by RandomizedSearchCV.

Gamma (default=0): This parameter determines the minimum of the loss reduction for a subsequent partitioning leaf node. This parameter can range between 0 to 1, and for this study, an array of shapes [0.001,0.01,0.02,0.1,0.25,0.5,1] was selected for tunning by the RandomizedSearchCV.

It is noticeable that increasing lambda, alpha, and gamma results in more conservative models. Max_depth (default=3): Maximum depth determines the maximum number of allowable nodes from the root to the farthest leaf of a tree. Although the deeper trees by adding more nodes can reveal more complex relationships, they can cause overfitting of the model. This parameter was set for regression problem only, and a range of integer numbers between 1 to 20 was specified for tunning by the RandomizedSearchCV.

Table A - 5 XGBOOST regressor hyperparameters (full model)

Utility	**XGBOOST regressor hyperparameters (full model)**					
	Learning rate	**alpha**	**Min_child_weight**	**lambda**	**Max_Depth**	**Gamma**
Barrie	0.1	0.02	0.02	0.001	3	0.1
Calgary	0.5	0.02	0.001	0.02	15	0.001
Durham	0.5	0.02	0.001	0.02	15	0.001
Halifax	0.5	0.02	0.001	0.02	15	0.001
Kitchener	0.01	0.1	0.001	0.01	11	0.001
Markham	0.5	0.02	0.001	0.02	15	0.001
Region of Waterloo	0.3	0.01	0.01	0.1	5	0.01
Saskatoon	0.1	0.001	0.02	0.02	6	0.001
St.John’s	0.5	0.1	0.001	0.1	3	0.001
Vancouver	0.5	0.001	0.001	0.01	10	0.01
Victoria	0.5	0.01	0.1	0.01	7	0.001
Waterloo	0.5	0.02	0.001	0.02	15	0.001
Winnipeg	0.25	0.01	0.1	0.01	2	0.001

Table A - 6 XGBOOST regressor hyperparameters (model with selected features)

Utility	**XGBOOST regressor hyperparameters (model with selected features)**					
	Learning rate	**alpha**	**Min_child_weight**	**lambda**	**Max_Depth**	**Gamma**
Barrie	1	0.01	0.1	0.01	8	0.001
Calgary	0.5	0.02	0.001	0.02	15	0.001
Durham	0.5	0.02	0.001	0.02	15	0.001
Halifax	0.1	0.01	0.01	0.01	10	0.001
Kitchener	0.01	0.001	0.02	0.02	15	1

Markham	0.5	0.02	0.001	0.02	15	0.001
Region of Waterloo	0.3	0.01	0.01	0.1	5	0.01
Saskatoon	0.1	0.001	0.02	0.02	6	0.001
St.John's	0.5	0.1	0.001	0.1	3	0.001
Vancouver	0.5	0.02	0.001	0.02	15	0.001
Victoria	0.5	0.01	0.1	0.01	7	0.001
Waterloo	0.5	0.02	0.001	0.02	15	0.001
Winnipeg	0.25	0.01	0.1	0.01	2	0.001

Table A - 7 XGBOOST classifier hyperparameters (Full model)

Utility	**XGBOOST classifier hyperparameters (full model)**				
	eta	**alpha**	**Min_child_weight**	**lambda**	**Gamma**
Barrie	0.5	0.02	0.1	0.001	0.01
Calgary	0.1	0.1	0.02	0.001	0.02
Durham	0.1	0.1	0.02	0.001	0.02
Halifax	0.1	0.1	0.02	0.001	0.02
Kitchener	0.001	0.1	0.1	0.1	0.5
Markham	0.01	0.1	0.001	0.12	0.5
Region of Waterloo	0.001	0.1	0.1	0.1	0.5
Saskatoon	0.25	0.1	0.01	0.1	0.001
St.John's	0.25	0.1	0.01	0.01	0.01
Vancouver	0.3	0.01	0.1	0.001	0.35
Victoria	0.1	0.1	0.001	0.01	0.95
Waterloo	0.1	0.1	0.02	0.001	0.02
Winnipeg	0.1	0.1	0.02	0.001	0.02

Table A - 8XGBOOST classifier hyperparameters (model with selected features)

Utility	**XGBOOST classifier hyperparameters (model with selected features)**				
	eta	**alpha**	**Min_child_weight**	**lambda**	**Gamma**
Barrie	0.25	0.01	0.1	0.001	0.01
Calgary	0.1	0.01	0.02	0.02	0.001
Durham	1	0.01	0.1	0.001	0.1
Halifax	1	0.02	0.02	0.001	0.02
Kitchener	0.5	0.001	0.02	0.1	0.1

Markham	0.001	0.001	0.001	0.02	0.25
Region of Waterloo	0.5	0.001	0.02	0.1	0.1
Saskatoon	0.25	0.02	0.1	0.001	0.02
St.John's	0.25	0.01	0.1	0.001	0.01
Vancouver	0.95	0.001	0.1	0.1	0.45
Victoria	0.1	0.01	0.1	0.01	0.3
Waterloo	0.02	0.001	0.01	0.001	0.25
Winnipeg	0.25	0.02	0.1	0.001	0.02

Appendix B

B 1 Correlation analysis – Break Status

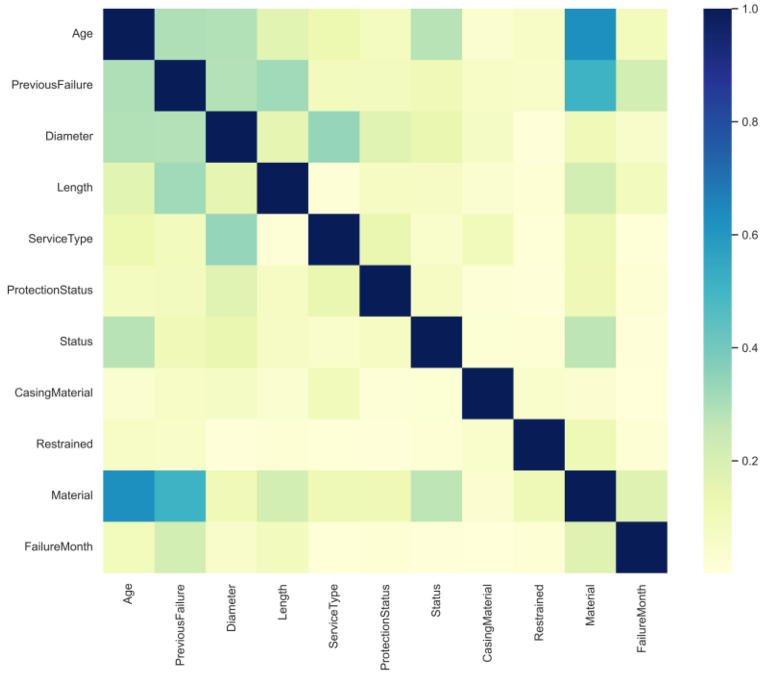


Figure B - 1 Correlation analysis - Break Status - Barrie

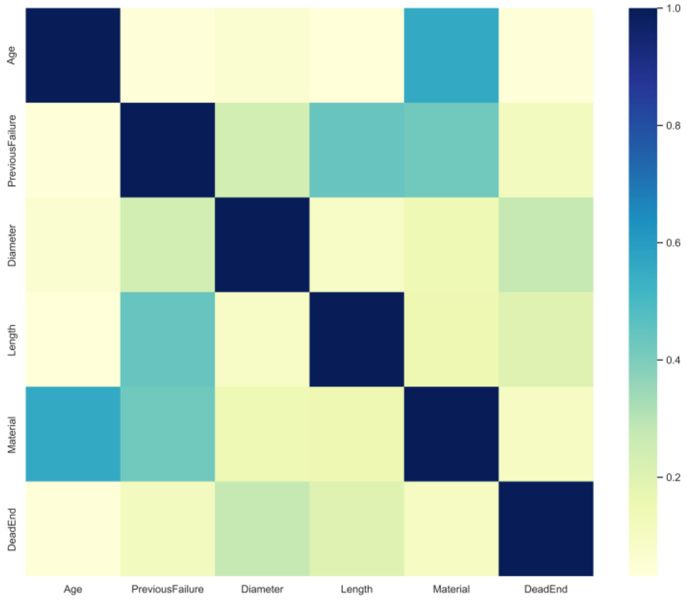


Figure B - 2 Correlation analysis - Break Status – Calgary

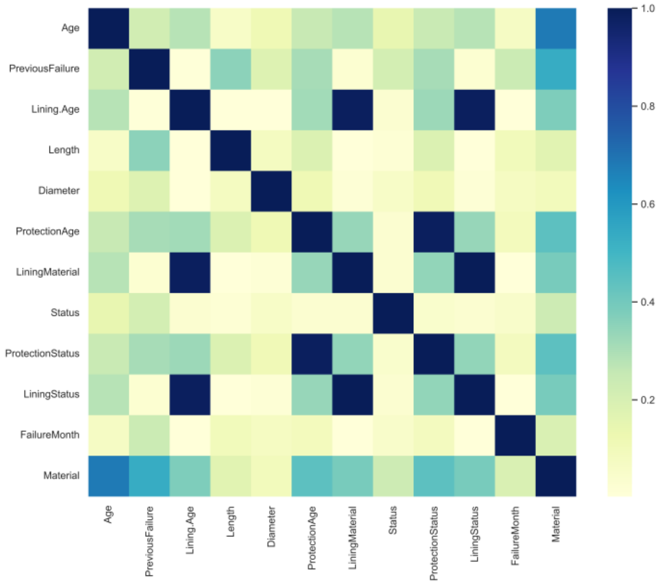


Figure B - 3 Correlation analysis - Break Status – Durham

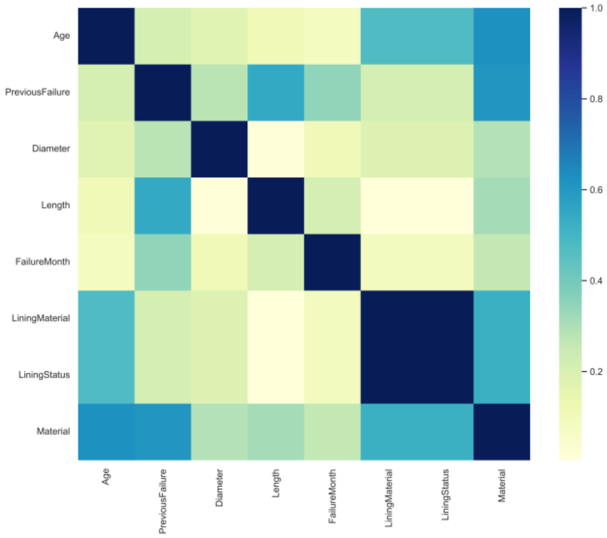


Figure B - 4 Correlation analysis - Break Status – Halifax

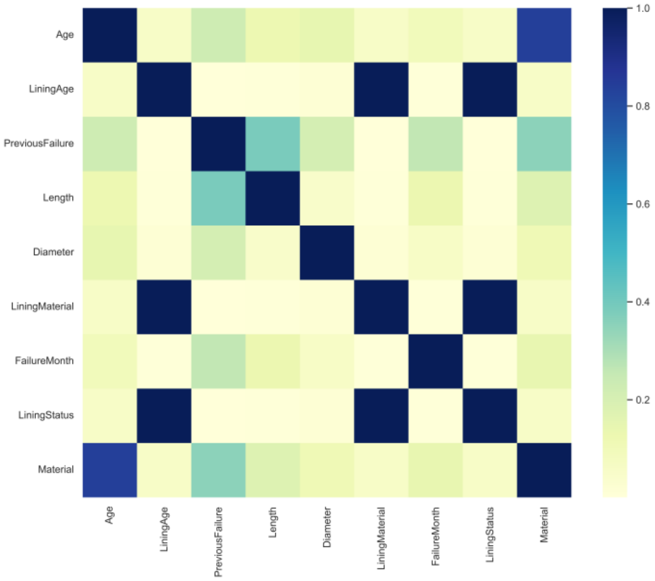


Figure B - 5 Correlation analysis - Break Status – Kitchener

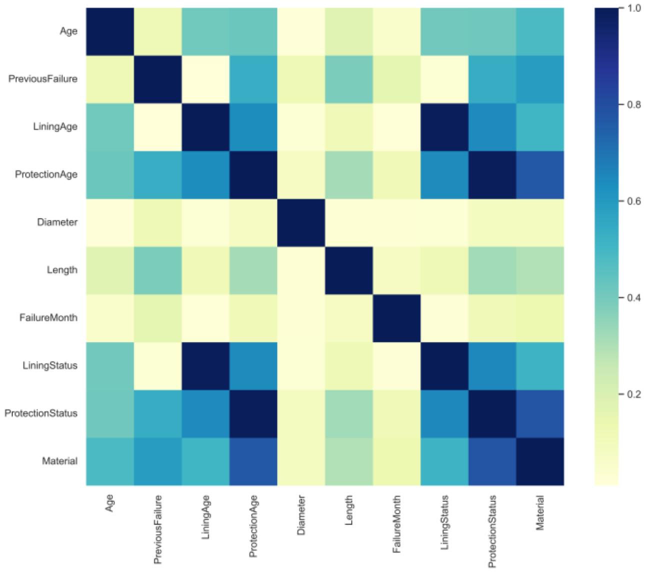


Figure B - 6 Correlation analysis - Break Status – Markham

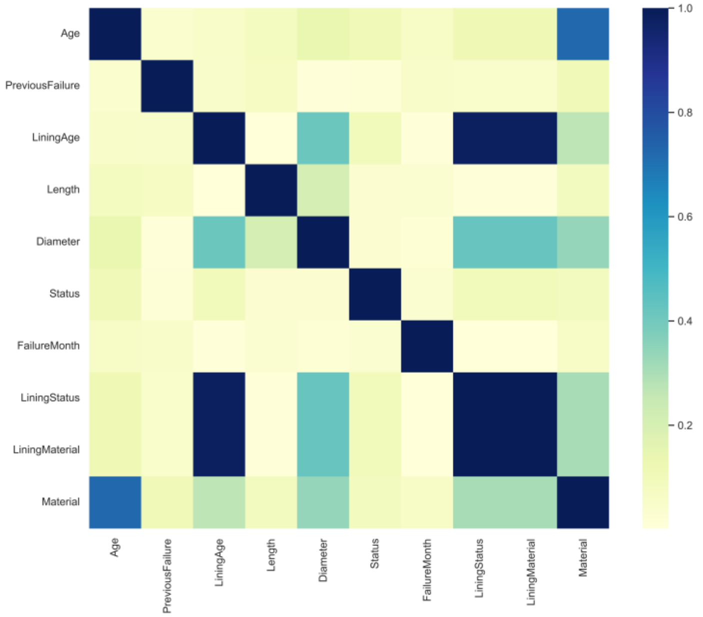


Figure B - 7 Correlation analysis - Break Status – RoW

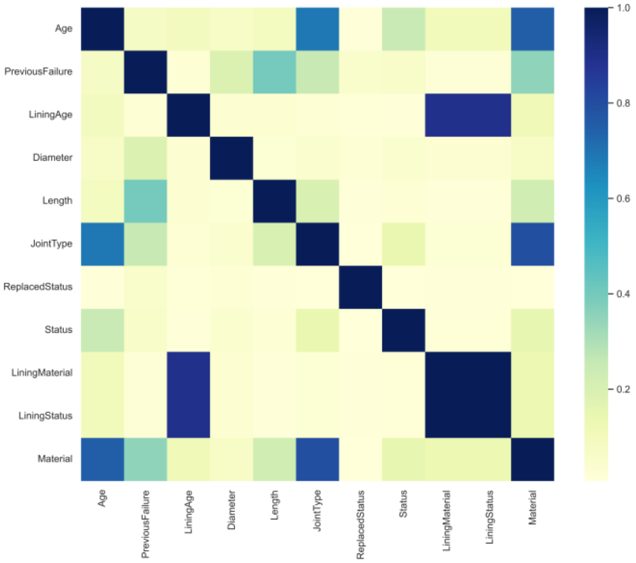


Figure B - 8 Correlation analysis - Break Status – Saskatoon

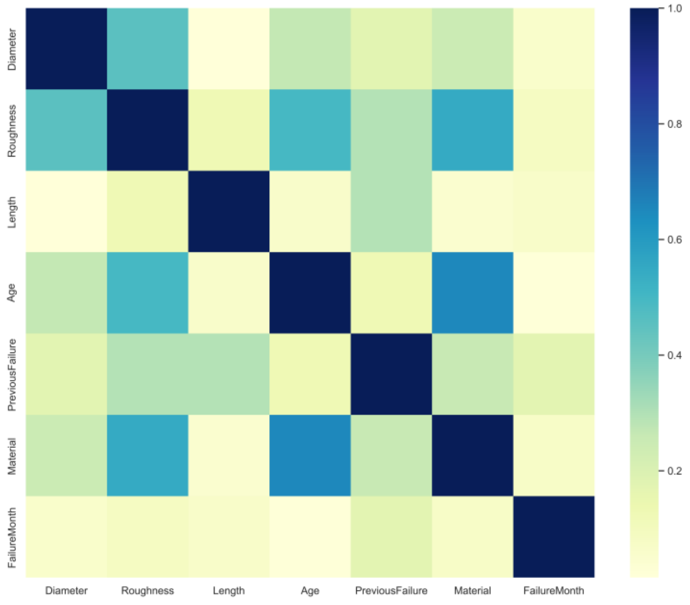


Figure B - 9 Correlation analysis - Break Status – St. Johns

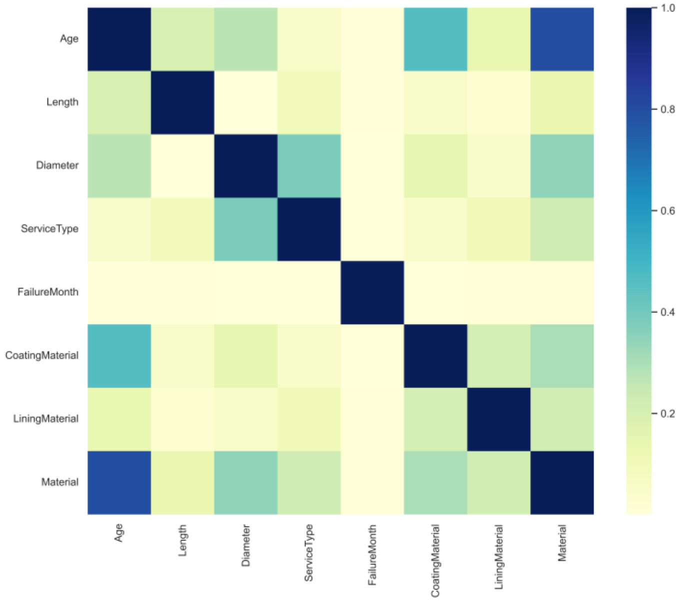


Figure B - 10 Correlation analysis - Break Status – Vancouver

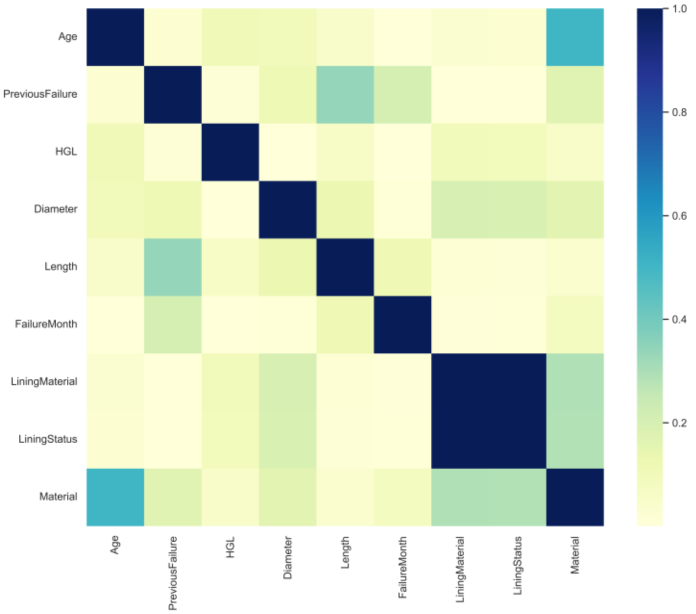


Figure B - 11 Correlation analysis - Break Status – Victoria

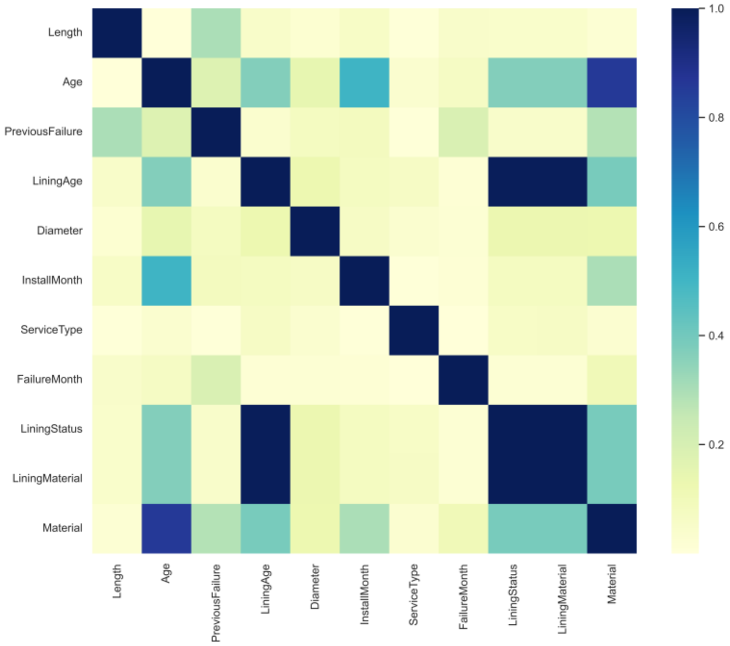


Figure B - 12 Correlation analysis - Break Status – Waterloo

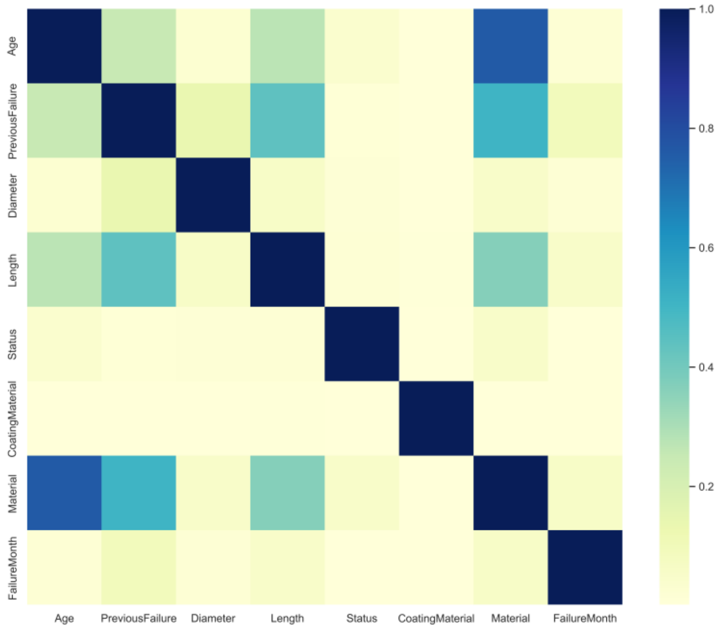


Figure B - 13 Correlation analysis - Break Status – Winnipeg

B 2 Correlation analysis – Current Rate of Failure

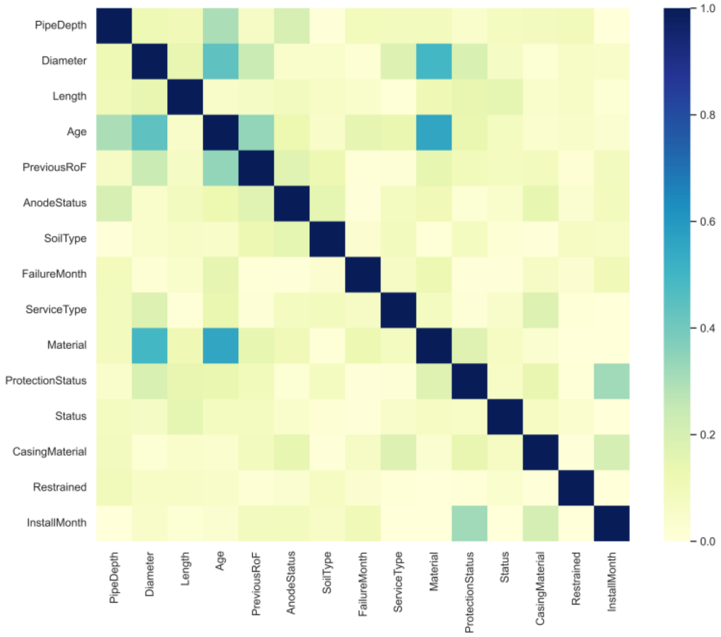


Figure B - 14 Correlation analysis - RoF – Barrie

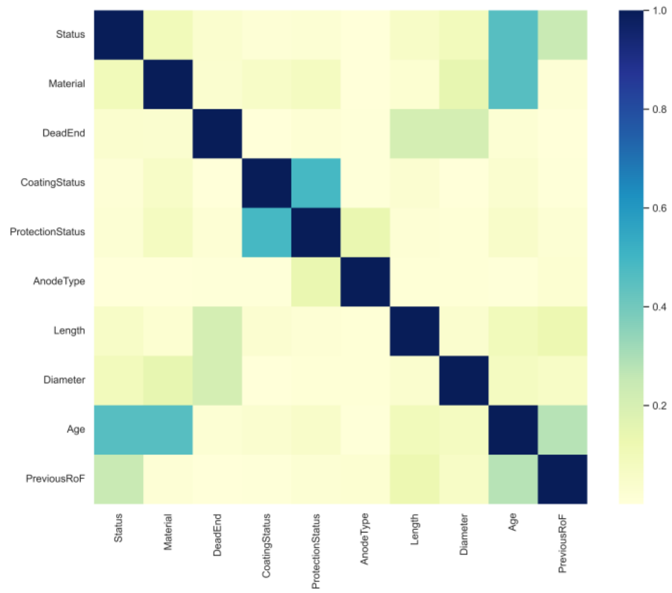


Figure B - 15 Correlation analysis - RoF – Calgary

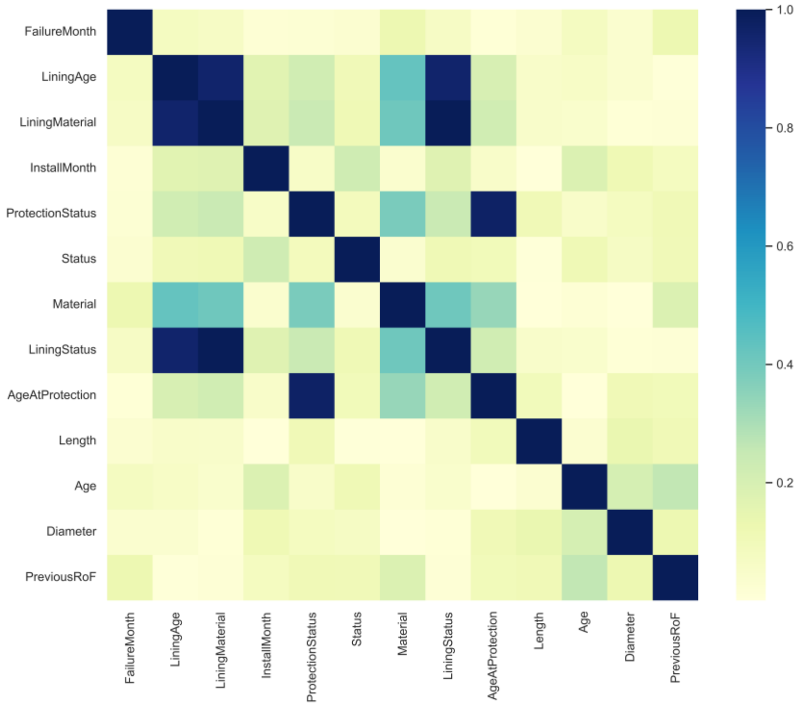


Figure B - 16 Correlation analysis - RoF – Durham

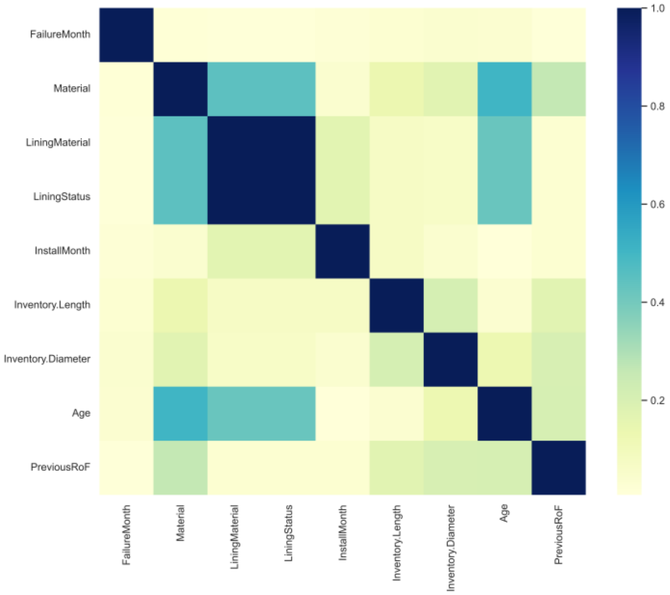


Figure B - 17 Correlation analysis - RoF – Halifax

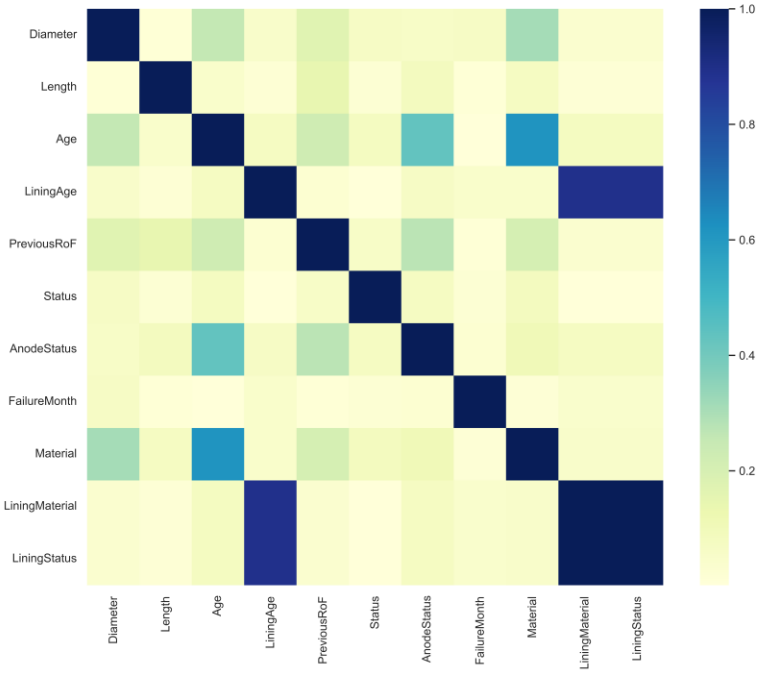


Figure B - 18 Correlation analysis - RoF – Kitchener

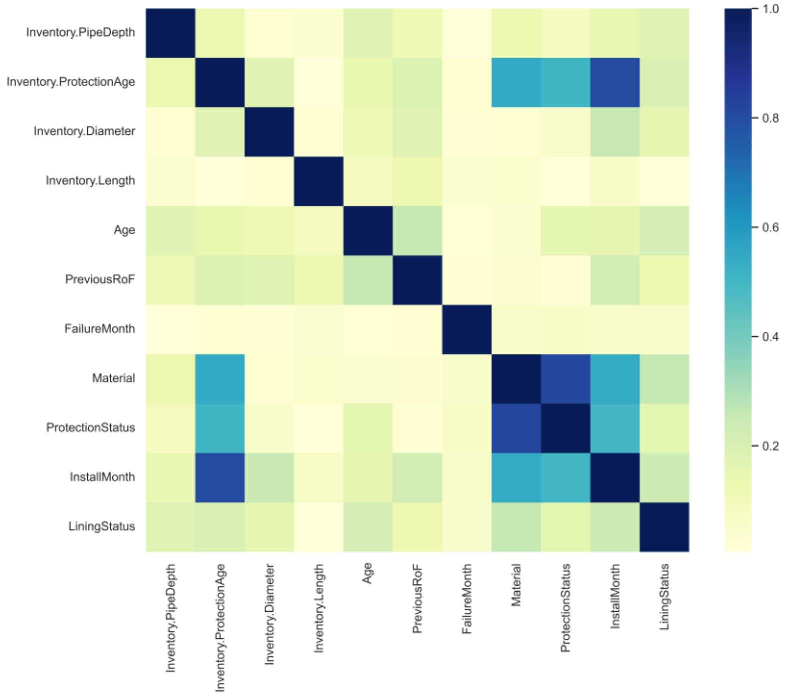


Figure B - 19 Correlation analysis - RoF – Markham

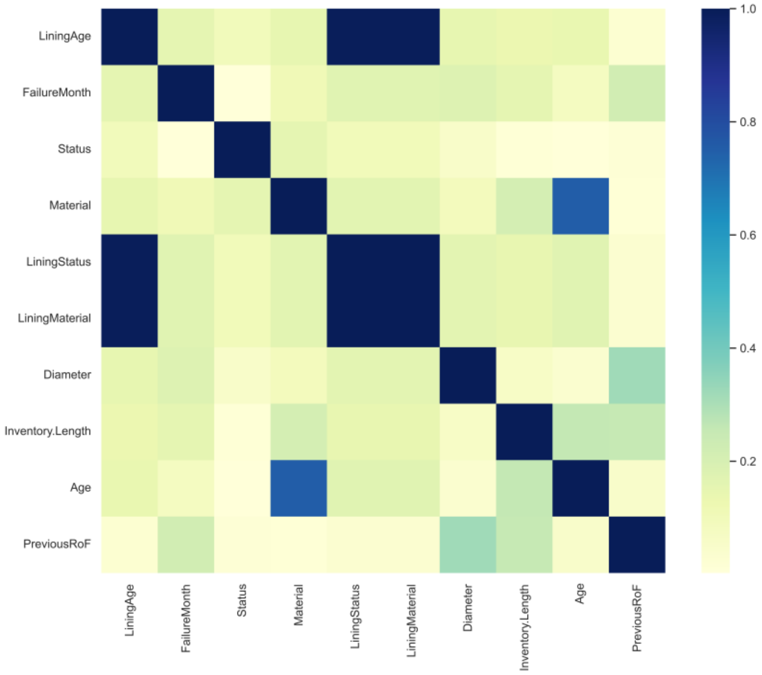


Figure B - 20 Correlation analysis - RoF – RoW

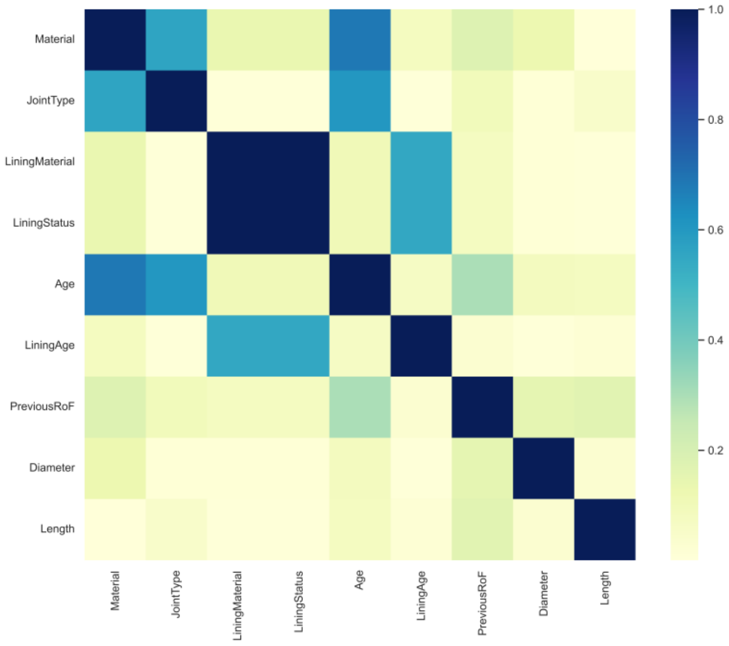


Figure B - 21 Correlation analysis - RoF – Saskatoon

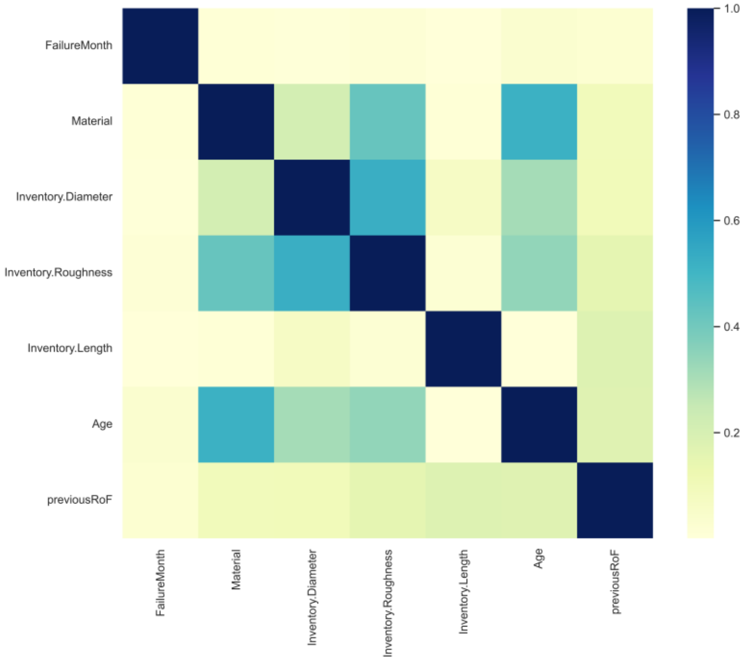


Figure B - 22 Correlation analysis - RoF – St.Johns

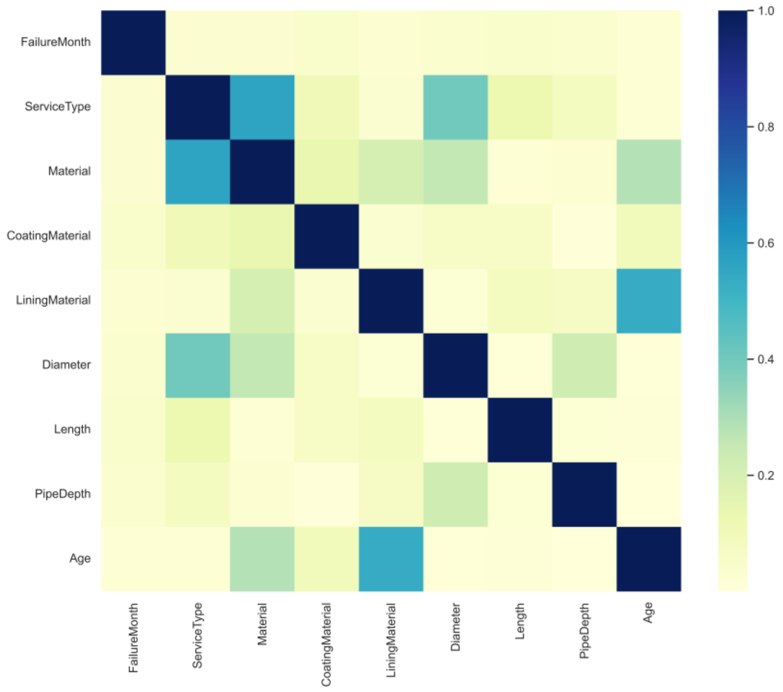


Figure B - 23 Correlation analysis - RoF – Vancouver

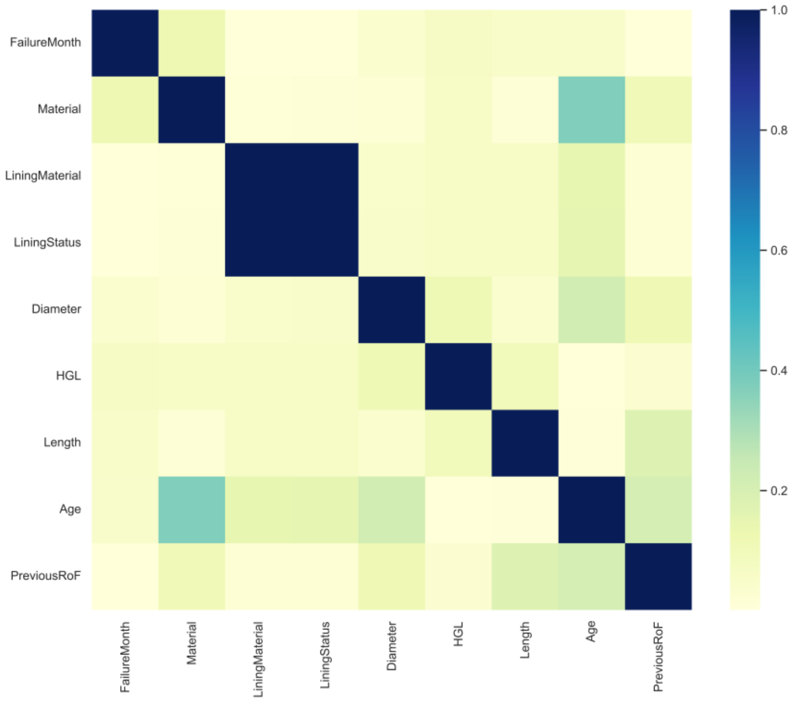


Figure B - 24 Correlation analysis - RoF – Victoria

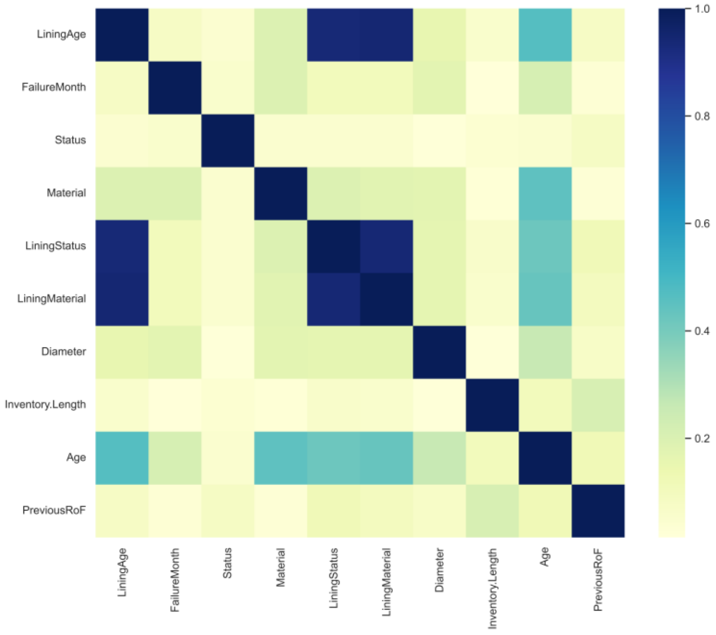


Figure B - 25 Correlation analysis - RoF – Waterloo

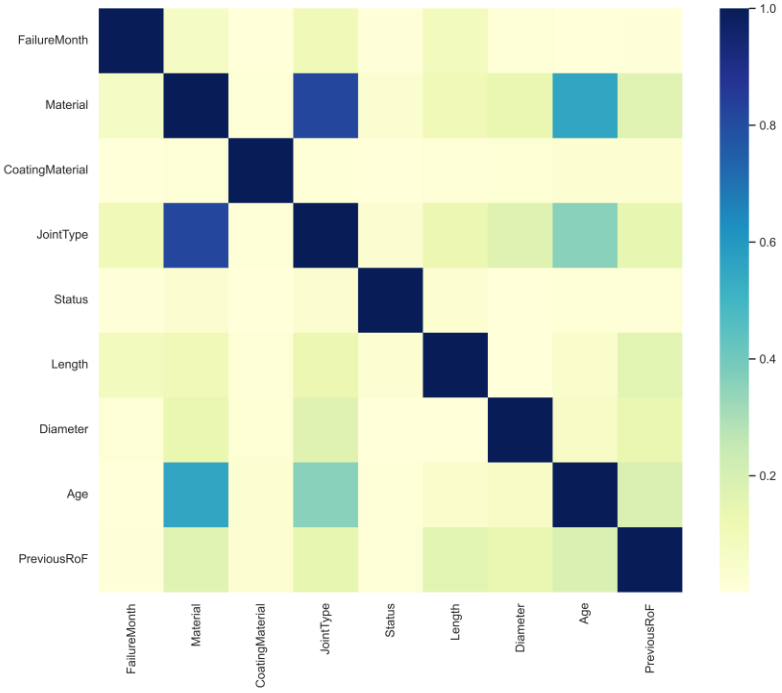


Figure B - 26 Correlation analysis - RoF – Winnipeg

Appendix C

C 1 R code for FAMD

```
library(readr)

data<-read.csv("C:\\Users\\SADAF2021\\Desktop\\Data-RoF\\Barrie\\Barrie-RoF-All.csv")

# Convert all columns to factor

df <- as.data.frame(unclass(data),stringsAsFactors = TRUE)

str(df)

library(FactoMineR)

res.famd <- FAMD(df,ncp=15, sup.var=16, graph = FALSE)

print(res.famd)

library("factoextra")

eig.val <- get_eigenvalue(res.famd)

eig.val

fviz_screeplot(res.famd)

var <- get_famd_var(res.famd)

var

coordinate <-data.frame(var$coord)

coordinate

library(writexl)

write_xlsx(coordinate,"C:\\Users\\SADAF2021\\Desktop\\CurrentRate-of-Failure\\Barrie\\FAMD\\Barrie-Coordinate.xlsx")

library("factoextra")

eig.val <- get_eigenvalue(res.famd)

eig.val
```

```
variance <-data.frame(eig.val)

library(writexl)

write_xlsx(variance,"C:\\Users\\SADAF2021\\Desktop\\CurrentRate-
of-Failure\\Barrie\\FAMD\\Barrie-variance.xlsx")

var$cos2

var$contrib

contribution <- data.frame(var$contrib)

contribution

library(writexl)

write_xlsx(contribution,"C:\\Users\\SADAF2021\\Desktop\\CurrentR
ate-of-Failure\\Barrie\\FAMD\\Barrie-contribution.xlsx")

fviz_contrib(res.famd,"var", axes=1:15)

fviz_contrib(res.famd,"var", axes=2)

fviz_cos2(res.famd,"var",axes=1:15)
```

C 1 R code for CATPCA

```
install.packages("Gifi", repos="http://R-Forge.R-project.org")

library("Gifi")

library(factoextra)

data<-            read.csv("C:\\Users\\SADAF2021\\Desktop\\Data-
RoF\\Barrie\\Barrie-RoF-All.csv")

# Convert all columns to factor

data3<-as.data.frame(unclass(data),
stringsAsFactors = TRUE)

str(data3)

ActiveVariables<-
c(TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,TR
UE,TRUE,TRUE,FALSE)
```

```
DegreeVec <-c(-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,1,1,1,1,1,1)

CATPCA <- princals(data3, ndim = 15, ordinal = FALSE, ties = "s",
knots = knotsGifi(data3, "E"),degrees = DegreeVec, copies = 1,
missing = "m", normobj.z = TRUE, active = ActiveVariables,itmax =
1000, eps = 1e-06, verbose = FALSE)

summary(CATPCA)

EigenVector <-CATPCA$loadings

EigenVector

EigenValue <-CATPCA$evals

Eigenvector<-data.frame(EigenVector)

Eigenvalue <- data.frame(EigenValue)

CATPCA$scoremat

library(writexl)

write_xlsx(Eigenvector,"C:\\Users\\SADAF2021\\Desktop\\CurrentRa
te-of-Failure\\Barrie\\CATPCA\\Barrie-Eigenvector.xlsx")

write_xlsx(Eigenvalue,
"C:\\Users\\SADAF2021\\Desktop\\CurrentRate-of-
Failure\\Barrie\\CATPCA\\Barrie-Eigenvalue.xlsx")

CATPCA$evals

CATPCA$dmeasures

plot(CATPCA,"screeplot")
```

C 2 R code for Optimal scaling

```
install.packages("Gifi", repos="http://R-Forge.R-project.org")

library(optiscale)

library(tidyverse)

library(caret)

library(leaps)
library(MASS)

library(writexl)

data<-      read.csv("C:\\Users\\SADAF2021\\Desktop\\BreakStatus-FinalCSV\\BreakStatus-Transformed\\Barrie\\Barrie                -BreakStatus.csv")

data3 <- as.data.frame(unclass(data),                  # Convert all columns to factor

                       stringsAsFactors = TRUE)

str(data3)

for (i in 1:8 ) {



  Qualitive <-data3[,i]

  op.scaled<-
opscale(x.qual=Qualitive,x.quant=seq(1:length(Qualitive)),level=1, process=1)

data3 <- data.frame(data3,op.scaled$os)}

data3

data3 <- data3[ -c(1,1:8) ]

data3

write_xlsx(data3,"C:\\Users\\SADAF2021\\Desktop\\BreakStatus-FinalCSV\\BreakStatus-Transformed\\Barrie-Transformed-BS.xlsx")
```

C 3 Python code for RF-RFECV (Current rate of failure)

```
# In this note book the following steps are taken:
1. Remove highly correlated attributes
2. Find the best hyper parameters for estimator(RF)
3. Check fitting of the full model
4. Calculate r2 of the full model
5. Find the most important features by tunned XGBOOST
6. Find the best hyper parameter of the model with selected features
7. Comapring r2 of the tuuned full model and model with selected features

import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.feature_selection import RFECV
from sklearn.model_selection import train_test_split, GridSearchCV, KFold,
RandomizedSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn import metrics
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import accuracy_score,r2_score
import xgboost
from xgboost import XGBRFRegressor, XGBRegressor
from sklearn.metrics import make_scorer
r2_score = make_scorer(r2_score)

## Barrie

#import data
Data=pd.read_csv("Barrie-Transfomed-Data.csv")

X = Data.iloc[:,:-1]
y = Data.iloc[:,-1]

#split test and training set. total number of data is 330 so the test size
cannot be large
np.random.seed(60)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.09,
                                                    random_state = 1000)

np.random.seed(60)
regressors = {}
regressors.update({"Random Forest":
RandomForestRegressor(random_state=1000)})

#Define range of hyperparameters for estimator
np.random.seed(60)
parameters = {}
parameters.update({"Random Forest": {
                                    "regressor__n_estimators":
[10,20,30,40,50,60,70,80,90,100,110,120,130,140,150,160,170,180,190,200],
                                    "regressor__max_features": ["auto",
"sqrt", "log2"],
                                    "regressor__max_depth" :
[5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20],
                                    #"regressor__min_samples_split": [2, 5,
10,15],
                                    #"regressor__min_samples_leaf":
[1,2,4,6],
                                    "regressor__bootstrap":[True,False]
}})

# Make correlation matrix
corr_matrix = X_train.corr(method = "spearman").abs()

# Draw the heatmap
sns.set(font_scale = 1.0)
f, ax = plt.subplots(figsize=(11, 9))
sns.heatmap(corr_matrix, cmap= "YlGnBu", square=True, ax = ax)
f.tight_layout()
plt.savefig("Barrie-RoF-correlation_matrix.png", dpi = 1080)

# Select upper triangle of matrix
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k =
1).astype(np.bool))

# Find index of feature columns with correlation greater than 0.8
to_drop = [column for column in upper.columns if any(upper[column] > 0.8)]

# Drop features
X_train = X_train.drop(to_drop, axis = 1)
X_test = X_test.drop(to_drop, axis = 1)

X_train

FEATURE_IMPORTANCE = {"Random Forest"}

np.random.seed(60)
selected_regressor = "Random Forest"
regressor = regressors[selected_regressor]

np.random.seed(60)
scaler = StandardScaler()
steps = [("scaler", scaler), ("regressor", regressor)]
pipeline = Pipeline(steps = steps)

#Define parameters that we want to use in gridsearch cv
param_grid = parameters[selected_regressor]

# Initialize GridSearch object for estimator
gscv = RandomizedSearchCV(pipeline, param_grid, cv = 3,  n_jobs= -1, verbose
= 1, scoring = r2_score, n_iter=30)

np.random.seed(60)
results = {}
for regressor_label, regressor in regressors.items():
    # Print message to user
    print(f"Now tuning {regressor_label}.")

# Fit gscv (Tunes estimator)
```

```
print(f"Now tuning {selected_regressor}. Go grab a beer or something.")
gscv.fit(X_train, np.ravel(y_train))

#Getting the best hyperparameters
best_params = gscv.best_params_
best_params

#Getting the best score of model
best_score = gscv.best_score_
best_score

#Check overfitting of the estimator
from sklearn.model_selection import cross_val_score
mod = RandomForestRegressor(
 max_depth= 16,
 max_features= 'auto',
    bootstrap=False,
 n_estimators= 30 ,random_state=10000)

scores_test = cross_val_score(mod, X_test, y_test, scoring='r2', cv=5)

scores_test

tuned_params = {item[11:]: best_params[item] for item in best_params}
regressor.set_params(**tuned_params)

#Find r2 score of the model with all features (Model is tuned for all
features)
results={}
model=regressor.set_params( max_depth= 16,
 max_features= 'auto',
    bootstrap=False,
 n_estimators= 30 ,random_state=10000)
model.fit(X_train,y_train)
y_pred = model.predict(X_test)
R2 = metrics.r2_score(y_test, y_pred)
results = {"regressor": model,
              "Best Parameters": best_params,
              "Training r2": best_score*100,
              "Test r2": R2*100}
results


# Select Features using RFECV
class PipelineRFE(Pipeline):
    # Source: https://ramhiser.com/post/2018-03-25-feature-selection-with-
scikit-learn-pipeline/
    def fit(self, X, y=None, **fit_params):
        super(PipelineRFE, self).fit(X, y, **fit_params)
        self.feature_importances_ = self.steps[-1][-1].feature_importances_
        return self

steps = [("scaler", scaler), ("regressor", regressor)]
pipe = PipelineRFE(steps = steps)
np.random.seed(60)

# Initialize RFECV object
feature_selector = RFECV(pipe, cv = 5, step = 1, verbose = 1)

# Fit RFECV
feature_selector.fit(X_train, np.ravel(y_train))

# Get selected features
feature_names = X_train.columns
selected_features = feature_names[feature_selector.support_].tolist()

performance_curve = {"Number of Features": list(range(1, len(feature_names) +
1)),
                    "R2": feature_selector.grid_scores_}
performance_curve = pd.DataFrame(performance_curve)

# Performance vs Number of Features
# Set graph style
sns.set(font_scale = 1.75)
sns.set_style({"axes.facecolor": "1.0", "axes.edgecolor": "0.85",
"grid.color": "0.85",
               "grid.linestyle": "-", 'axes.labelcolor': '0.4',
"xtick.color": "0.4",
               'ytick.color': '0.4'})
colors = sns.color_palette("RdYlGn", 20)
line_color = colors[3]
marker_colors = colors[-1]

# Plot
f, ax = plt.subplots(figsize=(13, 6.5))
sns.lineplot(x = "Number of Features", y = "R2", data = performance_curve,
             color = line_color, lw = 4, ax = ax)
sns.regplot(x = performance_curve["Number of Features"], y =
performance_curve["R2"],
            color = marker_colors, fit_reg = False, scatter_kws = {"s": 200},
ax = ax)

# Axes limits
plt.xlim(0.5, len(feature_names)+0.5)
plt.ylim(0.60, 1)

# Generate a bolded horizontal line at y = 0
ax.axhline(y = 0.625, color = 'black', linewidth = 1.3, alpha = .7)

# Turn frame off
```

```
ax.set_frame_on(False)

# Tight layout
plt.tight_layout()

#Define new training and test set based based on selected features by RFECV
X_train_rfecv = X_train[selected_features]
X_test_rfecv= X_test[selected_features]

np.random.seed(60)
regressor.fit(X_train_rfecv, np.ravel(y_train))

#Finding important features
np.random.seed(60)
feature_importance = pd.DataFrame(selected_features, columns = ["Feature
Label"])
feature_importance["Feature Importance"] = regressor.feature_importances_
feature_importance = feature_importance.sort_values(by="Feature Importance",
ascending=False)
feature_importance

# Initialize GridSearch object for model with selected features
np.random.seed(60)
gscv = RandomizedSearchCV(pipeline, param_grid, cv = 3,  n_jobs= -1, verbose
= 1, scoring = r2_score, n_iter=30)

#Tuning random forest REGRESSOR with selected features
np.random.seed(60)
gscv.fit(X_train_rfecv,y_train)

#Getting the best parameters of model with selected features
best_params = gscv.best_params_
best_params

#Getting the score of model with selected features
best_score = gscv.best_score_
best_score

#Check overfitting of the  tuned model with selected features
from sklearn.model_selection import cross_val_score
mod = RandomForestRegressor(max_depth= 9,
 max_features='auto' ,
bootstrap=False,
 n_estimators= 180 ,random_state=10000)

scores_test = cross_val_score(mod, X_test_rfecv, y_test, scoring='r2', cv=5)

scores_test

results={}
```

```
model=regressor.set_params(max_depth= 9,
 max_features='auto' ,
bootstrap=False,
 n_estimators= 180 ,random_state=10000)
model.fit(X_train_rfecv,y_train)
y_pred = model.predict(X_test_rfecv)
R2 = metrics.r2_score(y_test, y_pred)
results = {"regressorr": model,
              "Best Parameters": best_params,
              "Training r2": best_score*100,
              "Test r2": R2*100}
results
```

C 4 Python codes for XGBOOST-RFECV (Current rate of failure)

```
# In this note book the following steps are taken:
1. Find the best hyper parameters for estimator(XGBOOST)
2. Check fitting of the full model
3. Calculate r2 of the full model
4. Find the most important features by tunned XGBOOST
5. Find the best hyper parameter of the model with selected features
6. Comapring r2 of the tuuned full model and model with selected features

import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.feature_selection import RFECV
from sklearn.model_selection import train_test_split, GridSearchCV, KFold,
RandomizedSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn import metrics
from sklearn.metrics import accuracy_score,r2_score
import xgboost
from xgboost import XGBRegressor
from sklearn.metrics import make_scorer
r2_score = make_scorer(r2_score)

## Barrie

#import data
Data=pd.read_csv("Barrie-Transfomed-Data.csv")

X = Data.iloc[:,:-1]
y = Data.iloc[:,-1]
```

```
#split test and training set. total number of data is 330 so the test size
cannot be large
np.random.seed(60)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
                                                    random_state = 1000)

np.random.seed(60)
regressors = {}
regressors.update({"XGBoost": XGBRegressor(random_state=1000)})

#Define range of hyperparameters for estimator
np.random.seed(60)
parameters = {}
parameters.update({"XGBoost": {

"regressor__learning_rate":[0.001,0.01,0.02,0.1,0.25,0.5,1],

"regressor__gamma":[0.001,0.01,0.02,0.1,0.25,0.5,1],
                                    "regressor__max_depth" :
[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20],

"regressor__reg_alpha":[0.001,0.01,0.02,0.1],

"regressor__reg_lambda":[0.001,0.01,0.02,0.1],

"regressor__min_child_weight":[0.001,0.01,0.02,0.1]
}})

# Make correlation matrix
corr_matrix = X_train.corr(method = "spearman").abs()

# Draw the heatmap
sns.set(font_scale = 1.0)
f, ax = plt.subplots(figsize=(11, 9))
sns.heatmap(corr_matrix, cmap= "YlGnBu", square=True, ax = ax)
f.tight_layout()
plt.savefig("correlation_matrix.png", dpi = 1080)

# Select upper triangle of matrix
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k =
1).astype(np.bool))

# Find index of feature columns with correlation greater than 0.8
to_drop = [column for column in upper.columns if any(upper[column] > 0.8)]

# Drop features
X_train = X_train.drop(to_drop, axis = 1)
X_test = X_test.drop(to_drop, axis = 1)
```

```
X_train

FEATURE_IMPORTANCE = {"XGBoost"}

np.random.seed(60)
selected_regressor = "XGBoost"
regressor = regressors[selected_regressor]

np.random.seed(60)
scaler = StandardScaler()
steps = [("scaler", scaler), ("regressor", regressor)]
pipeline = Pipeline(steps = steps)

#Define parameters that we want to use in gridsearch cv
param_grid = parameters[selected_regressor]

# Initialize GridSearch object for estimator
gscv = RandomizedSearchCV(pipeline, param_grid, cv = 3,  n_jobs= -1, verbose
= 1, scoring = r2_score, n_iter=40)

np.random.seed(60)
results = {}
for regressor_label, regressor in regressors.items():
    # Print message to user
    print(f"Now tuning {regressor_label}.")

# Fit gscv (Tunes estimator)
print(f"Now tuning {selected_regressor}. Go grab a beer or something.")
gscv.fit(X_train, np.ravel(y_train))

#Getting the best hyperparameters
best_params = gscv.best_params_
best_params

#Getting the best score of model
best_score = gscv.best_score_
best_score

#Check overfitting of the estimator
from sklearn.model_selection import cross_val_score
mod = XGBRegressor(reg_lambda=0.02,
                           reg_alpha=0.02,
                           min_child_weight=0.1,
                           max_depth=10,
                           learning_rate=0.25,
                           gamma=0.001
 ,random_state=10000)

scores_test = cross_val_score(mod, X_test, y_test, scoring='r2', cv=5)
```

```
scores_test

tuned_params = {item[11:]: best_params[item] for item in best_params}
regressor.set_params(**tuned_params)

#Find r2 score of the model with all features (Model is tuned for all 
features)
results={}
model=regressor.set_params(reg_lambda=0.01,
                            reg_alpha=0.01,
                            min_child_weight=0.01,
                            max_depth=8,
                            learning_rate=1,
                            gamma=0.001
 ,random_state=10000)
model.fit(X_train,y_train)
y_pred = model.predict(X_test)
R2 = metrics.r2_score(y_test, y_pred)
results = {"regressor": model,
              "Best Parameters": best_params,
              "Training r2": best_score*100,
              "Test r2": R2*100}
results


# Select Features using RFECV
class PipelineRFE(Pipeline):
    # Source: https://ramhiser.com/post/2018-03-25-feature-selection-with-
scikit-learn-pipeline/
    def fit(self, X, y=None, **fit_params):
        super(PipelineRFE, self).fit(X, y, **fit_params)
        self.feature_importances_ = self.steps[-1][-1].feature_importances_
        return self

steps = [("scaler", scaler), ("regressor", regressor)]
pipe = PipelineRFE(steps = steps)
np.random.seed(60)

# Initialize RFECV object
feature_selector = RFECV(pipe, cv = 5, step = 1, verbose = 1)

# Fit RFECV
feature_selector.fit(X_train, np.ravel(y_train))

# Get selected features
feature_names = X_train.columns
selected_features = feature_names[feature_selector.support_].tolist()

performance_curve = {"Number of Features": list(range(1, len(feature_names) + 
1)),
                    "R2": feature_selector.grid_scores_}
performance_curve = pd.DataFrame(performance_curve)

# Performance vs Number of Features
# Set graph style
sns.set(font_scale = 1.75)
sns.set_style({"axes.facecolor": "1.0", "axes.edgecolor": "0.85",
"grid.color": "0.85",
               "grid.linestyle": "-", 'axes.labelcolor': '0.4',
"xtick.color": "0.4",
               'ytick.color': '0.4'})
colors = sns.color_palette("RdYlGn", 20)
line_color = colors[3]
marker_colors = colors[-1]

# Plot
f, ax = plt.subplots(figsize=(13, 6.5))
sns.lineplot(x = "Number of Features", y = "R2", data = performance_curve,
             color = line_color, lw = 4, ax = ax)
sns.regplot(x = performance_curve["Number of Features"], y =
performance_curve["R2"],
            color = marker_colors, fit_reg = False, scatter_kws = {"s": 200},
ax = ax)

# Axes limits
plt.xlim(0.5, len(feature_names)+0.5)
plt.ylim(0.60, 1)

# Generate a bolded horizontal line at y = 0
ax.axhline(y = 0.625, color = 'black', linewidth = 1.3, alpha = .7)

# Turn frame off
ax.set_frame_on(False)

# Tight layout
plt.tight_layout()

#Define new training and test set based based on selected features by RFECV
X_train_rfecv = X_train[selected_features]
X_test_rfecv= X_test[selected_features]

np.random.seed(60)
regressor.fit(X_train_rfecv, np.ravel(y_train))

#Finding important features
np.random.seed(60)
feature_importance = pd.DataFrame(selected_features, columns = ["Feature
Label"])
feature_importance["Feature Importance"] = regressor.feature_importances_
```

```
feature_importance = feature_importance.sort_values(by="Feature Importance",
ascending=False)
feature_importance

# Initialize GridSearch object for model with selected features
np.random.seed(60)
gscv = RandomizedSearchCV(pipeline, param_grid, cv = 3,  n_jobs= -1, verbose
= 1, scoring = r2_score, n_iter=30)

#Tuning random forest REGRESSOR with selected features
np.random.seed(60)
gscv.fit(X_train_rfecv,y_train)

#Getting the best parameters of model with selected features
best_params = gscv.best_params_
best_params

#Getting the score of model with selected features
best_score = gscv.best_score_
best_score

#Check overfitting of the  tuned model with selected features
from sklearn.model_selection import cross_val_score
mod = XGBRegressor(reg_lambda=0.01,
                            reg_alpha=0.01,
                            min_child_weight=0.1,
                            max_depth=8,
                            learning_rate=1,
                            gamma=0.001
 ,random_state=10000)

scores_test = cross_val_score(mod, X_test_rfecv, y_test, scoring='r2', cv=5)

scores_test

results={}
model=regressor.set_params(reg_lambda=0.01,
                            reg_alpha=0.01,
                            min_child_weight=0.1,
                            max_depth=8,
                            learning_rate=1,
                            gamma=0.001
 ,random_state=10000)
model.fit(X_train_rfecv,y_train)
y_pred = model.predict(X_test_rfecv)
R2 = metrics.r2_score(y_test, y_pred)
results = {"regressorr": model,
              "Best Parameters": best_params,
              "Training r2": best_score*100,
              "Test r2": R2*100}
```

```
results
```

C 5 Python code for RF-RFECV (Break Status)

```
# In this note book the following steps are taken:

1. Remove highly correlated attributes

2. Find the best hyper parameters for estimator

3. Find the most important features by tunned random forest

4. Find f1 score of the tunned full model

5. Find best hyper parameter of model with selected features

6. Find f1 score of the tuned seleccted model

7. Compare the two f1 scores


import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.feature_selection import RFECV,RFE

from sklearn.model_selection import train_test_split, GridSearchCV, KFold,RandomizedSearchCV

from sklearn.preprocessing import StandardScaler

from sklearn.pipeline import Pipeline

from sklearn import metrics

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score,f1_score,recall_score

import numpy as np

from sklearn.metrics import make_scorer

f1_score = make_scorer(f1_score)
```

```
Recall=make_scorer(recall_score)


#import data
Data=pd.read_csv("RandomForest-Data/Barrie-Transformed-BS.csv")


X = Data.iloc[:,:-1]
y = Data.iloc[:,-1]


#split test and training set.
np.random.seed(60)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.1,
                                                    random_state = 1000)


#Define estimator and model
classifiers = {}
classifiers.update({"Random                                          Forest":
RandomForestClassifier(random_state=1000)})


#Define range of hyperparameters for estimator
np.random.seed(60)
parameters = {}
parameters.update({"Random     Forest":     {     "classifier__n_estimators":
[100,105,110,115,120,125,130,135,140,145,150,155,160,170,180,190,200],

                                   #              "classifier__n_estimators":
[2,4,5,6,7,8,9,10,20,30,40,50,60,70,80,90,100,110,120,130,140,150,160,170,180
,190,200],

                                    #"classifier__class_weight":        [None,
"balanced"],
                                    "classifier__max_features":       ["auto",
"sqrt", "log2"],

                                     "classifier__max_depth"                  :
[4,6,8,10,11,12,13,14,15,16,17,18,19,20,22],

                                    #"classifier__max_depth"                  :
[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20],

                                    "classifier__criterion"          :["gini",
"entropy"]


}})


# Make correlation matrix

corr_matrix = X_train.corr(method = "spearman").abs()


# Draw the heatmap

sns.set(font_scale = 1.0)

f, ax = plt.subplots(figsize=(11, 9))

sns.heatmap(corr_matrix, cmap= "YlGnBu", square=True, ax = ax)

f.tight_layout()

plt.savefig("Barrie_Rf_correlation_matrix.png", dpi = 1080)


# Select upper triangle of matrix

upper    =    corr_matrix.where(np.triu(np.ones(corr_matrix.shape),    k    =
1).astype(np.bool))


# Find index of feature columns with correlation greater than 0.8

to_drop = [column for column in upper.columns if any(upper[column] > 0.8)]
```

```
# Drop features
X_train = X_train.drop(to_drop, axis = 1)
X_test = X_test.drop(to_drop, axis = 1)

X_train

FEATURE_IMPORTANCE = {"Random Forest"}
selected_classifier = "Random Forest"
classifier = classifiers[selected_classifier]

scaler = StandardScaler()
steps = [("scaler", scaler), ("classifier", classifier)]
pipeline = Pipeline(steps = steps)

#Define parameters that we want to use in gridsearch cv
param_grid = parameters[selected_classifier]

# Initialize GridSearch object for estimator
gscv = RandomizedSearchCV(pipeline, param_grid, cv = 3,  n_jobs= -1, verbose =
1, scoring = f1_score, n_iter=50)

# Fit gscv (Tunes estimator)
print(f"Now tuning {selected_classifier}. Go grab a beer or something.")
gscv.fit(X_train, np.ravel(y_train))

#Getting the best hyperparameters
```

```
best_params = gscv.best_params_

best_params


#Getting the best score of model

best_score = gscv.best_score_

best_score


#Check overfitting of the estimator

from sklearn.model_selection import cross_val_score

mod = RandomForestClassifier(criterion= 'entropy',

 max_depth= 19,

 max_features= 'log2',

 n_estimators= 180 ,random_state=10000)


scores_test = cross_val_score(mod, X_test, y_test, scoring='f1', cv=5)


scores_test


tuned_params = {item[12:]: best_params[item] for item in best_params}

classifier.set_params(**tuned_params)


#Find f1 score of the model with all features (Model is tuned for all features)

results={}

model=classifier.set_params(criterion= 'entropy',

 max_depth= 19,

 max_features= 'log2',
 n_estimators= 180 ,random_state=10000)
model.fit(X_train,y_train)
y_pred = model.predict(X_test)
F1 = metrics.f1_score(y_test, y_pred)
Recall=recall_score(y_test, y_pred)
results = {"classifier": model,
              "Best Parameters": best_params,
              "Training f1": best_score*100,
              "Test f1": F1*100, "Test recall": Recall*100}
results


# Select Features using RFECV
class PipelineRFE(Pipeline):
    #   Source:   https://ramhiser.com/post/2018-03-25-feature-selection-with-scikit-learn-pipeline/
    def fit(self, X, y=None, **fit_params):
        super(PipelineRFE, self).fit(X, y, **fit_params)
        self.feature_importances_ = self.steps[-1][-1].feature_importances_
        return self

steps = [("scaler", scaler), ("classifier", classifier)]
pipe = PipelineRFE(steps = steps)
np.random.seed(60)

# Initialize RFECV object
```

```
feature_selector = RFECV(pipe, cv = 5, step = 1, verbose = 1)


# Fit RFECV

feature_selector.fit(X_train, np.ravel(y_train))


# Get selected features

feature_names = X_train.columns

selected_features = feature_names[feature_selector.support_].tolist()


selected_features


performance_curve = {"Number of Features": list(range(1, len(feature_names) +
1)),

                     "F1": feature_selector.grid_scores_}

performance_curve = pd.DataFrame(performance_curve)


# Performance vs Number of Features

# Set graph style

sns.set(font_scale = 1.75)

sns.set_style({"axes.facecolor": "1.0", "axes.edgecolor": "0.85",
"grid.color": "0.85",

               "grid.linestyle": "-", 'axes.labelcolor': '0.4', "xtick.color":
"0.4",

               'ytick.color': '0.4'})

colors = sns.color_palette("RdYlGn", 20)

line_color = colors[3]

marker_colors = colors[-1]
```

```
# Plot

f, ax = plt.subplots(figsize=(13, 6.5))

sns.lineplot(x = "Number of Features", y = "F1", data = performance_curve,

             color = line_color, lw = 4, ax = ax)

sns.regplot(x = performance_curve["Number of Features"], y =
performance_curve["F1"],

            color = marker_colors, fit_reg = False, scatter_kws = {"s": 200},
ax = ax)


# Axes limits

plt.xlim(0.5, len(feature_names)+0.5)

plt.ylim(0.60, 1)


# Generate a bolded horizontal line at y = 0

ax.axhline(y = 0.625, color = 'black', linewidth = 1.3, alpha = .7)


# Turn frame off

ax.set_frame_on(False)


# Tight layout

plt.tight_layout()


#Define new training and test set based based on selected features by RFECV

X_train_rfecv = X_train[selected_features]

X_test_rfecv= X_test[selected_features]
```

```
np.random.seed(60)
classifier.fit(X_train_rfecv, np.ravel(y_train))


#Finding important features
np.random.seed(60)
feature_importance = pd.DataFrame(selected_features, columns = ["Feature Label"])
feature_importance["Feature Importance"] = classifier.feature_importances_
feature_importance = feature_importance.sort_values(by="Feature Importance", ascending=False)
feature_importance


# Initialize GridSearch object for model with selected features
np.random.seed(60)
gscv = RandomizedSearchCV(pipeline, param_grid, cv = 3,  n_jobs= -1, verbose = 1, scoring = f1_score, n_iter=30)


#Tuning random forest classifier with selected features
np.random.seed(60)
gscv.fit(X_train_rfecv,y_train)


#Getting the best parameters of model with selected features
best_params = gscv.best_params_
best_params


#Getting the score of model with selected features
best_score = gscv.best_score_
```

```
best_score

#Check overfitting of the  tuned model with selected features
from sklearn.model_selection import cross_val_score
mod = RandomForestClassifier(#class_weight= None,
 criterion= 'gini',
 max_depth= 18,
 max_features= 'log2',
 n_estimators= 100 ,random_state=10000)

scores_test = cross_val_score(mod, X_test_rfecv, y_test, scoring='f1', cv=5)

scores_test

results={}
model=classifier.set_params(criterion= 'gini',
 max_depth= 18,
 max_features= 'log2',
 n_estimators= 100 ,random_state=10000)
model.fit(X_train_rfecv,y_train)
y_pred = model.predict(X_test_rfecv)
F1 = metrics.f1_score(y_test, y_pred)
Recall=recall_score(y_test, y_pred)
results = {"classifier": model,
              "Best Parameters": best_params,
              "Training f1": best_score*100,
              "Test f1": F1*100,

          "Test recall": Recall*100}

results
```

C 6 Python Code for XGBOOST-RFECV (Break Status)

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.feature_selection import RFECV,RFE
from sklearn.model_selection import train_test_split, GridSearchCV,
KFold,RandomizedSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn import metrics
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score,f1_score
import numpy as np
from sklearn.metrics import make_scorer
f1_score = make_scorer(f1_score)
Recall=make_scorer(recall_score)

#import data
Data=pd.read_csv("XGBOOST-Data/Barrie-Transformed-BS-XG.csv")

X = Data.iloc[:,:-1]
y = Data.iloc[:,-1]

#split test and training set.
np.random.seed(60)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
                                                    random_state = 1000)

#Define estimator and model
classifiers = {}
classifiers.update({"XGBoost":
XGBClassifier(random_state=1000,eval_metric=f1_score,use_label_encoder=False)
})

#Define range of hyperparameters for estimator
np.random.seed(60)
parameters = {}
parameters.update({"XGBoost":
{"classifier__eta":[0.001,0.01,0.02,0.1,0.25,0.5,1],

"classifier__alpha":[0.001,0.01,0.02,0.1],
                                     "classifier__min_child_weight" :
[0.001,0.01,0.02,0.1],
                                    "classifier__lambda"
:[0.001,0.01,0.02,0.1],
                                     "classifier__gamma"
:[0.001,0.01,0.02,0.1,0.25,0.5,1],
                                     #"classifier__max_depth":
[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,1920]

}})


# Make correlation matrix
corr_matrix = X_train.corr(method = "spearman").abs()

# Draw the heatmap
sns.set(font_scale = 1.0)
f, ax = plt.subplots(figsize=(11, 9))
sns.heatmap(corr_matrix, cmap= "YlGnBu", square=True, ax = ax)
f.tight_layout()
plt.savefig("Barrie_XG_correlation_matrix.png", dpi = 1080)

# Select upper triangle of matrix
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k =
1).astype(np.bool))

# Find index of feature columns with correlation greater than 0.8
to_drop = [column for column in upper.columns if any(upper[column] > 0.8)]

# Drop features
X_train = X_train.drop(to_drop, axis = 1)
X_test = X_test.drop(to_drop, axis = 1)

X_train

FEATURE_IMPORTANCE = {"XGBoost"}
selected_classifier = "XGBoost"
classifier = classifiers[selected_classifier]

scaler = StandardScaler()
steps = [("scaler", scaler), ("classifier", classifier)]
pipeline = Pipeline(steps = steps)

#Define parameters that we want to use in gridsearch cv
param_grid = parameters[selected_classifier]

# Initialize gridsearchCV object for estimator
gscv =RandomizedSearchCV(pipeline, param_grid, cv = 3, n_jobs=-1, verbose =
3, scoring = f1_score, n_iter =10)
```

```
# Fit gscv (Tunes estimator)
print(f"Now tuning {selected_classifier}. Go grab a beer or something.")
gscv.fit(X_train, np.ravel(y_train))

#Getting the best hyperparameters
best_params = gscv.best_params_
best_params

#Getting the best score of model
best_score = gscv.best_score_
best_score

#Check overfitting of the estimator
from sklearn.model_selection import cross_val_score
mod = XGBClassifier(alpha=0.02,
                    eta= 0.5,
                    gamma= 0.01,
                    reg_lambda=0.001,
                     #max_Depth=7,
                    min_child_weight=0.1,
                    eval_metric='mlogloss',
                    random_state=10000)

scores_test = cross_val_score(mod, X_test, y_test, scoring='f1', cv=5)

scores_test

tuned_params = {item[12:]: best_params[item] for item in best_params}
classifier.set_params(**tuned_params)

#Find f1 score of the model with all features (Model is tuned for all
features)
results={}
model=classifier.set_params(alpha=0.02,
                    eta= 0.5,
                    gamma= 0.01,
                    reg_lambda=0.001,
                     #max_Depth=7,
                    min_child_weight=0.1,
                    eval_metric='mlogloss',
                    random_state=10000)

model.fit(X_train,y_train)
y_pred = model.predict(X_test)
F1 = metrics.f1_score(y_test, y_pred)
Recal=recall_score(y_test, y_pred)
results = {"classifier": model,
              "Best Parameters": best_params,
              "Training f1": best_score*100,
              "Test f1": F1*100, "Test recall":Recal*100}
```

```
results

# Select Features using RFECV
class PipelineRFE(Pipeline):
    # Source: https://ramhiser.com/post/2018-03-25-feature-selection-with-
scikit-learn-pipeline/
    def fit(self, X, y=None, **fit_params):
        super(PipelineRFE, self).fit(X, y, **fit_params)
        self.feature_importances_ = self.steps[-1][-1].feature_importances_
        return self

steps = [("scaler", scaler), ("classifier", classifier)]
pipe = PipelineRFE(steps = steps)
np.random.seed(60)

# Initialize RFECV object
feature_selector = RFECV(pipe, cv = 5, step = 1, verbose = 3)

# Fit RFECV
feature_selector.fit(X_train, np.ravel(y_train))

# Get selected features
feature_names = X_train.columns
selected_features = feature_names[feature_selector.support_].tolist()

performance_curve = {"Number of Features": list(range(1, len(feature_names) +
1)),
                    "F1": feature_selector.grid_scores_}
performance_curve = pd.DataFrame(performance_curve)

# Performance vs Number of Features
# Set graph style
sns.set(font_scale = 1.75)
sns.set_style({"axes.facecolor": "1.0", "axes.edgecolor": "0.85",
"grid.color": "0.85",
               "grid.linestyle": "-", 'axes.labelcolor': '0.4',
"xtick.color": "0.4",
               'ytick.color': '0.4'})
colors = sns.color_palette("RdYlGn", 20)
line_color = colors[3]
marker_colors = colors[-1]

# Plot
f, ax = plt.subplots(figsize=(13, 6.5))
sns.lineplot(x = "Number of Features", y = "F1", data = performance_curve,
             color = line_color, lw = 4, ax = ax)
sns.regplot(x = performance_curve["Number of Features"], y =
performance_curve["F1"],
            color = marker_colors, fit_reg = False, scatter_kws = {"s": 200},
ax = ax)
```

```
# Axes limits
plt.xlim(0.5, len(feature_names)+0.5)
plt.ylim(0.60, 1)

# Generate a bolded horizontal line at y = 0
ax.axhline(y = 0.625, color = 'black', linewidth = 1.3, alpha = .7)

# Turn frame off
ax.set_frame_on(False)

# Tight layout
plt.tight_layout()

#Define new training and test set based based on selected features by RFECV
X_train_rfecv = X_train[selected_features]
X_test_rfecv= X_test[selected_features]

np.random.seed(60)
classifier.fit(X_train_rfecv, np.ravel(y_train))

#Finding important features
np.random.seed(60)
feature_importance = pd.DataFrame(selected_features, columns = ["Feature
Label"])
feature_importance["Feature Importance"] = classifier.feature_importances_
feature_importance = feature_importance.sort_values(by="Feature Importance",
ascending=False)
feature_importance

# Initialize GridSearch object for model with selected features
np.random.seed(60)
gscv = RandomizedSearchCV(pipeline, param_grid, cv = 3,  n_jobs= -1, verbose
= 3, scoring = f1_score, n_iter=50)

#Tuning random forest classifier with selected features
np.random.seed(60)
gscv.fit(X_train_rfecv,y_train)

#Getting the best parameters of model with selected features
best_params = gscv.best_params_
best_params

#Getting the score of model with selected features
best_score = gscv.best_score_
best_score

#Check overfitting of the  tuned model with selected features
from sklearn.model_selection import cross_val_score
mod = XGBClassifier(alpha=0.01,
                    eta=0.25,
                    gamma=0.01,
                    reg_lambda=0.001,
                    #max_depth=4,
                    min_child_weight=0.1,
                    eval_metric='mlogloss',
                    random_state=10000)

scores_test = cross_val_score(mod, X_test_rfecv, y_test, scoring='f1', cv=5)

scores_test

results={}
model=classifier.set_params(alpha=0.01,
                    eta=0.25,
                    gamma=0.01,
                    reg_lambda=0.001,
                    #max_depth=4,
                    min_child_weight=0.1,
                    eval_metric='mlogloss',
                    random_state=10000)
model.fit(X_train_rfecv,y_train)
y_pred = model.predict(X_test_rfecv)
F1 = metrics.f1_score(y_test, y_pred)
Recal=recall_score(y_test, y_pred)
results = {"classifier": model,
              "Best Parameters": best_params,
              "Training f1": best_score*100,
              "Test f1": F1*100,  "Test recall":Recal*100}
Results
```

Appendix D

D 1 Inventory data summery

Table D - 1 Characteristics of inventory data

Utilities	%CI Pipes	%DI Pipes	% PVC Pipes	% small diameter pipes <200 mm	% Dead-end	% Cathodic protected pipes	% Lined pipes	% Coated pipes	% Restrained pipes	% Active pipes	%Distribution pipes	% Replaced Pipes	Average age	Average Lining age	Average Pressure	Average Length (m)
Barrie	16	25	52	70	-	4	-	-	3	89	96	-	27	-	-	128
Calgary	21	23	47	57	14	-	-	-	-	-	-	-	30	-	-	110
Durham	17	20	54	69	-	13	13	-	-	94	-	-	31	32	-	129
Halifax	44	48	2	67	-	-	44	-	-	-	-	-	39	-	-	148
Kitchener	24	35	36	64	-	-	0.3	-	-	-	-	-	34	18	-	0.8
Markham	9	14	71	65	-	27	11	-	-	-	-	-	23	33	-	132
RoW	10	32	69	22	-	-	25	-	-	96	-	-	32	26	-	86
Saskatoon	19	0.2	68	73	-	-	0.9	-	-	97	-	0.2	34	5	-	38
St.John's	43	44	12	66	-	-	-	-	-	-	-	-	40	-	-	74
Vancouver	43	54	0.1	71	-	-	47	12	-	-	95	-	43	-	-	24
Victoria	49	37	7	75	-	-	5	-	-	-	-	-	56	-	83	108
Waterloo	31	15	53	72	-	-	11	-	-	-	99	-	36	14	-	62
Winnipeg	25	1	55	74	-	-	-	0.2	-	99	-	-	38	-	-	36

D 2 Break data summery

Table D - 2 Characteristics of broken data

Utilities	%CI Pipes	%DI Pipes	% PVC Pipes	% small diameter pipes <200 mm	% Dead-end	% Cathodic protected pipes	% Lined pipes	% Coated pipes	% Anodic protection	% Restrained pipes	% Active pipes	%Distribution pipes	Average age	Average Lining age	Average Pressure	Average Length (m)	Average pipe depth (m)
Barrie	58	30	7	85	-	3	-	-	33	1	82	99	27	-	-	185	1.9
Calgary	49	36	9	72	8	1	-	0.3	1	-	93	-	30	-	-	151	-
Durham	38	33	18	77	-	32	21	-	-	-	90	-	31	38	-	180	-
Halifax	74	20	3	78	-	-	33	-	-	-	-	-	39		-	200	-
Kitchener	70	27	2	83	-	-	1	-	54	-	99	-	34	0.08	-	1.7	-
Markham	32	53	11	63	-	72	52	-	-	-	-	-	23	-	-	217	0.87
RoW	36	40	8	13	-	-	12	-	-	-	95	-	32	4	-	163	-
Saskatoon	37	0.2	10	81	-	-	2	-	-	-	-	-	34	0.02	-	70	-
St.John's	82	15	1	77	-	-	-	-	-	-	-	-	40	-	-	105	-
Vancouver	88	4	0.1	87	-	-	28		-	-	-	96	43	-	-	43	1.4
Victoria	72	23	3	84	-	-	5	-	-	-	-	-	56	-	81	155	-
Waterloo	81	14	5	81	-	-	23	-	-	-	99	-	36	2	-	130	-
Winnipeg	65	2	4	80	-	-	-	0.1	-	-	99	-	38	-	-	67	-

Appendix E

E 1 The defined structure of the collected data

Table E - 1 Structure of the collected data

Attribute	Type	Categories	Unit	Format
Material	Categorical	Asbestos Cement	N.A	N.A
		Cast Iron		
		Concrete		
		Copper		
		Cross Linked Polyethylene		
		Ductile Iron		
		Galvanized Steel		
		HDPE		
		PVC		
		PVCB		
		PVCF		
		PVCO		
		Polybutylene		
		Polyethylene		
		Steel		
Joint type	Categorical	Bell and Spigot	N.A	N.A
		Collar		
		Flared end		
		Gasket		
		Grooved		
		Lead		
		Mechanical		
		Rubber		
		Threaded		
		Universal		
		Welded		
Failure/Install month	Categorical	January	N.A	N.A
		February		
		March		
		April		
		May		
		June		
		July		
		August		
		September		
		October		

		November		
		December		
Lining Material	Categorical	CIP	N.A	N.A
		CM		
		CoalTar		
		Epoxy		
		HDPE		
		Polyurea		
		UnLined		
Casing Material	Categorical	Concrete	N.A	N.A
		Polyethylene		
		Polystrene		
		Steel		
		StryFoam		
		Tunnel		
		NoCasing		
Coating Material	Categorical	Asbestos	N.A	N.A
		CoalTar		
		Concrete		
		Epoxy		
		Foam		
		FRC		
		PB		
		Polyethylene		
		StyroFoam		
		Urecon		
		Y-Jacket		
		Uncoated		
Anode Type	Categorical	Magnesium	N.A	N.A
		Zinc		
		NoAnode		
Soil Type	Categorical	Clay	N.A	N.A
		Granular		
		Gravel		
		Marsh		
		Mixed		
		Muck		
		Natural ground		
		Road base		
		Rock		
		Sand		
		Clay/Granular		
		Clay/Gravel		
		Clay/Muck		

		Clay/Rock		
		Clay/Silt		
		Clay/Stone		
		Clay/Stone/Peat		
		Granular/Rock		
		Gravel/Rock		
		Rock/Peat		
		Sand/Clay		
		Sand/Clay/Gravel		
		Sand/Clay/Loam		
		Sand/Clay/Rock		
		Sand/Clay/Till		
		Sand/Granular		
		Sand/Gravel		
		Sand/Muck		
		Sand/Peat		
		Sand/Rock		
		Sand/Silt		
		Sand/Stone		
		Stone/Concrete		
Service Type	Categorical	Distribution	N.A	N.A
		Facility		
		Service		
		Transmission		
Dead-end	Booliean	Yes No	N.A	N.A
Status				
Replaced Status				
Protection Status				
Lining Status				
Anode Status				
Diameter	Numerical	N.A	mm	N.A
Length	Numerical	N.A	m	N.A
Pipe Depth	Numerical	N.A	m	N.A
Installation/Failure Date	Date	N.A	N.A	YYYY/MM/D