EXAMPLE Flow monitor data are obtained from a given sewer basin, and dry weather flow conditions are characterized for both Weekday and Weekend Day Groups. Weekday Q_{min}, Q_{avg}, and Q_{max} are 0.103 MGD, 0.428 MGD, and 0.726 MGD, respectively. Weekend Q_{min}, Q_{avg}, and Q_{max} are 0.109 MGD, 0.459 MGD, and 0.873 MGD, respectively.

Use the % Minimum Method to estimate Q_{GWI} and Q_{WW}. Assume that x = 80%.

(a) Calculate Q_{GWI} for the Weekday Day Group using Equation (2).

$Q_{GWI} = xQ_{min} = (0.80)(0.103) = 0.082$ MGD

(b) Do the same for the Weekend Day Group. $Q_{GWI} = 0.087$ MGD

(c) Identify the Day Group with the minimum Q_{GWI}. This is the Weekday Day Group. Then, set Q_{GWI} equal to this value for both Day Groups. Therefore, Q_{GWI} is set equal to 0.082 MGD for both the Weekday and Weekend Day Groups.

(d) Calculate Q_{WW} for the Weekday Day Group using Equation (3).

$Q_{WW} = Q_{avg} - Q_{GWI} = 0.428 - 0.082 = 0.346$ MGD

(e) Do the same for the Weekend Day Group. $Q_{WW} = 0.377$ MGD

$$Q_{WW} = \frac{(Q_{avg} - Q_{min})}{x} \tag{4}$$

Values of 88% and 90% are commonly used, and values of 70% and 75% have also been proposed (DeCoite, 1981 and Mitchell, 2007). Once wastewater production is estimated using Equation (4), groundwater infiltration is estimated using Equation (5) by subtracting wastewater production from the average dry weather flow rate.

$$Q_{GWI} = Q_{avg} - Q_{WW} \tag{5}$$

Substituting Equation (4) into Equation (5) provides groundwater infiltration in one step as shown in Equation (6).

$$Q_{GWI} = Q_{avg} - \frac{(Q_{avg} - Q_{min})}{x} \tag{6}$$

The general assumption that $0 \leq Q_{GWI} \leq Q_{min}$ governs the allowable range of x, and this range is determined by solving Equation (6) for x when $Q_{GWI} = 0$ and $Q_{GWI} = Q_{min}$. Based on these assumptions, $1 - (Q_{min}/Q_{avg}) \leq x \leq 1$. The application of the Wastewater Production Method is demonstrated in the following example:

STEVENS-SCHUTZBACH METHOD

The *Stevens-Schutzbach Method* estimates groundwater infiltration based on the minimum and average flow rates using the relationship shown in Equation (7) (Mitchell, 2007).

EXAMPLE Flow monitor data are obtained from a given sewer basin, and dry weather flow conditions are characterized for both Weekday and Weekend Day Groups. Weekday Q_{min}, Q_{avg}, and Q_{max} are 0.103 MGD, 0.428 MGD, and 0.726 MGD, respectively. Weekend Q_{min}, Q_{avg}, and Q_{max} are 0.109 MGD, 0.459 MGD, and 0.873 MGD, respectively.

Use the Wastewater Production Method to estimate Q_{GWI} and Q_{WW}. Assume that x = 88%.

Solution

(a) Calculate Q_{GWI} for the Weekday Day Group using Equation (6).

$$Q_{GWI} = Q_{avg} - \frac{(Q_{avg} - Q_{min})}{x} = 0.428 - \frac{(0.428 - 0.103)}{0.88} = 0.059 \text{ MGD}$$

(b) Do the same for the Weekend Day Group. $Q_{GWI} = 0.061$ MGD

(e) Identify the Day Group with the minimum Q_{GWI}. This is the Weekday Day Group. Then, set Q_{GWI} equal to this value for both Day Groups. Therefore, Q_{GWI} is set equal to 0.059 MGD for both the Weekday and Weekend Day Groups.

(d) Calculate Q_{WW} for the Weekday Day Group using Equation (3).

$Q_{WW} = Q_{avg} - Q_{GWI} = 0.428 - 0.059 = 0.369$ MGD

Do the same for the Weekend Day Group. $Q_{WW} = 0.400$ MGD

$$Q_{GWI} = \frac{0.4Q_{min}}{1 - 0.6\left(\frac{Q_{min}}{Q_{avg}}\right)^{(Q_{avg})^{0.7}}} \quad (7)$$

Once groundwater infiltration is estimated using Equation (7), wastewater production is estimated using Equation (3) by subtracting groundwater infiltration from the average dry weather flow rate. The application of the Stevens-Schutzbach Method is demonstrated in the following example:

MITCHELL METHOD

The Mitchell Method uses the average flow rate and a minimum factor (MF) to determine an expected minimum flow (Q_{MF}). Actual and expected minimum flow rates are then used to estimate groundwater infiltration using an iterative solution as shown in Figure 2 (Mitchell, 2007).

The minimum factor is initially computed with Equation (8) assuming no groundwater, and groundwater infiltration is updated using Equation (9). Convergence is often achieved in no more than three iterations.

EXAMPLE Flow monitor data are obtained from a given sewer basin, and dry weather flow conditions are characterized for both Weekday and Weekend Day Groups. Weekday Q_{min}, Q_{avg}, and Q_{max} are 0.103 MGD, 0.428 MGD, and 0.726 MGD, respectively. Weekend Q_{min}, Q_{avg}, and Q_{max} are 0.109 MGD, 0.459 MGD, and 0.873 MGD, respectively.

Use the Stevens-Schutzbach Method to estimate Q_{GWI} and Q_{WW}.

Solution

(a) Calculate Q_{GWI} for the Weekday Day Group using Equation (7).

$$Q_{GWI} = \frac{0.4Q_{min}}{1 - 0.6\left(\frac{Q_{min}}{Q_{avg}}\right)^{(Q_{avg})^{0.7}}} = \frac{0.4(0.103)}{1 - 0.6\left(\frac{0.103}{0.428}\right)^{(0.428)^{0.7}}} = 0.057 \text{ MGD}$$

(b) Do the same for the Weekend Day Group. $Q_{GWI} = 0.059$ MGD

(c) Identify the Day Group with the minimum Q_{GWI}. This is the Weekday Day Group. Then, set Q_{GWI} equal to this value for both Day Groups. Therefore, Q_{GWI} is set equal to 0.057 MGD for both the Weekday and Weekend Day Groups.

(d) Calculate Q_{WW} for the Weekday Day Group using Equation (3).

$$Q_{WW} = Q_{avg} - Q_{GWI} = 0.428 - 0.057 = 0.371 \text{ MGD}$$

(e) Do the same for the Weekend Day Group. $Q_{WW} = 0.402$ MGD

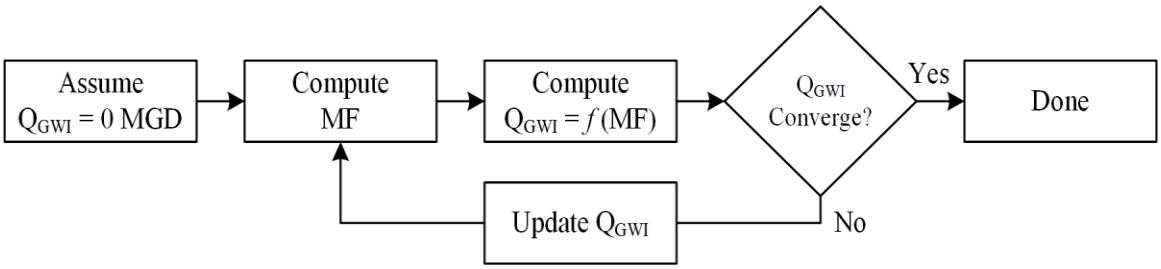

FIGURE 2: Iterative Solution for the Mitchell Method

$$MF = 0.222(Q_{avg} - Q_{GWI})^{0.202} \tag{8}$$

$$Q_{GWI} = \frac{Q_{min} - (MF)Q_{avg}}{1 - MF} \tag{9}$$

The minimum factor equation used here is the equation originally reported in the literature (Mitchell, 2007). However, other minimum factor equations have also been reported by various sources, and these equations are mathematically interchangeable. The application of the Mitchell Method is demonstrated in the following example:

EXAMPLE Flow monitor data are obtained from a given sewer basin, and dry weather flow conditions are characterized for both Weekday and Weekend Day Groups. Weekday Q_{min}, Q_{avg}, and Q_{max} are 0.103 MGD, 0.428 MGD, and 0.726 MGD, respectively. Weekend Q_{min}, Q_{avg}, and Q_{max} are 0.109 MGD, 0.459 MGD, and 0.873 MGD, respectively.

Use the Mitchell Method to estimate Q_{GWI} and Q_{WW}.

Solution

(a) Calculate Q_{BI} for the Weekday Day Group.

Apply iterative procedure shown in Figure 2.

			assume			calculate
Iteration	Q_{min}	Q_{avg}	Q_{GWI}	MF	$Q_{min\ MF}$	Q_{GWI}
	MGD	MGD	MGD		MGD	MGD
1	0.103	0.428	0.000	0.19	0.081	0.028
2	0.103	0.428	0.028	0.18	0.076	0.033
3	0.103	0.428	0.033	0.18	0.076	0.033

(b) Do the same for the Weekend Day Group. Q_{GWI} = 0.032 MGD

(c) Identify the Day Group with the minimum Q_{GWI}. This is the Weekend Day Group. Then, set Q_{GWI} equal to this value for both Day Groups. Therefore, Q_{GWI} is set equal to 0.032 MGD for both the Weekday and Weekend Day Groups.

(d) Calculate Q_{WW} for the Weekday Day Group using Equation (3).

$Q_{WW} = Q_{avg} - Q_{GWI} = 0.428 - 0.032 = 0.396$ MGD

Do the same for the Weekend Day Group. Q_{WW} = 0.427 MGD

A NOTE ABOUT UNITS

When applying these four methods, it is important to understand implications regarding units of measure. The % Minimum and the Wastewater Production Method can be used with various flow rate units of measure, provided that consistent units of measure are used. However, both the Stevens-Schutzbach Method and the Mitchell Method require that units of million gallons per day (MGD) be used for the minimum and average flow rate, and the resulting wastewater production and groundwater infiltration estimates are provided in the same units, as well.

CONCLUSION

The previous sections demonstrate *how* to use various methods to estimate groundwater infiltration in sewers using flow monitor data. However, they do not demonstrate *when* to select one method over another. The order in which these methods are presented reflects the historical timeline on which they were developed from the % Minimum Method and Wastewater Production Method in the 1970s to the Stevens-Schutzbach Method in the 1990s to the Mitchell Method in the 2000s, and the newer methods accommodate or resolve weaknesses of the older

methods. Neither the % Minimum Method nor the Wastewater Production Method account for basin size. As basin size increases, flow attenuation occurs. This phenomenon is not acknowledged by these two methods, and as a result, they tend to overstate groundwater infiltration in larger basins. Both the Stevens-Schutzbach Method and the Mitchell Method accommodate flow attenuation and provide more realistic estimates of groundwater infiltration in larger basins. The Stevens-Schutzbach Method does this entirely on empirical grounds, while the Mitchell Method accomplishes this while rooted in established minimum factors familiar from sewer design. This paper provides details for wastewater professionals to use each of these methods, as well as appropriate details and caveats to consider along the way.

ACKNOWLEDGEMENT

The authors acknowledge Ralph Petroff, Jim Schutzbach, and Paul Mitchell for their knowledge and expertise regarding groundwater infiltration at its estimation from sewer flow monitor data. They have provided many helpful discussions regarding the development and proper applications of these methods.

REFERENCES

Enfinger, K.L. and Stevens, P.L. (2006). "Sewer Sociology – The Days of Our (Sewer) Lives." Proceedings of the Water Environment Federation Technical Exhibition and Conference; Dallas, TX; Water Environment Federation: Alexandria, VA.

Vallabhaneni, S., Chan, C.C., and Burgess, E.H. (2007). Computer Tools for Sanitary Sewer System Capacity Planning, United States Environmental Protection Agency, Office of Research and Development: Washington, D.C. EPA/600/R-07/111.

DeCoite, D.C.W, Tsugita, R.A., and Petroff, R. (1981). "Infiltration/Inflow Source Identification by Comprehensive Flow Monitoring," Journal of the Water Pollution Control Federation, Volume 53, Issue 11, 1620-1626.

Mitchell, Stevens, and Nazaroff. (2007). A Comparison of Methods and a Simple Empirical Solution to Quantifying Base Infiltration in Sewers. Water Practice, Volume 1, No. 6, 2007 Water Environment Federation.

Evaluation of Artificial Intelligence Tool Performance for Predicting Water Pipe Failures

Sepideh Yazdekhasti, Ph.D.[1]; Greta Vladeanu, Ph.D.[2]; and Craig Daly, P.E.[3]

[1]Research Analyst, Columbia, MD. Email: Sepideh.yazdekhasti@xyleminc.com
[2]Engineering Analyst, Columbia, MD. Email: Greta.Vladeanu@xyleminc.com
[3]Senior Program Manager, Columbia, MD. Email: Craig.Daly@xyleminc.com

ABSTRACT

Over the past years, there has been a sustained interest in developing machine learning (ML) models that are sophisticated enough to capture the failure trends of water distribution systems and that are able to predict future breaks of the pipeline system. Given the limited budgetary resources of water pipeline owners, coupled with the deteriorated state of water networks, there is a vital need to deploy such tools to prioritize inspection and replacement of vulnerable regions, as well as to mitigate the chance of having catastrophic failures within the system. This study extends several ML algorithms that analyze the historical failures of water pipelines, with the goal to predict future breaks. The performance of each algorithm has been examined using various water networks as different case studies with varying network size and configurations. The developed models are all aimed to estimate the future likelihood of pipe failure by exploring historical failure patterns, surrounding attributes (e.g., environmental and demographic), as well as pipe characteristics. To improve the predictive power of the learning algorithms, several engineered features have been also created from raw data and tested to facilitate the learning process of each algorithm. While developing such models is by no means an insignificant task, an equally, if not more important emphasis should be put on how precisely these models are predicting actual failures. Additionally, the model variables should be defined wisely enough to ensure the uniqueness of each network has been captured and incorporated into the analysis. Lastly, it is crucial to evaluate the precision of the developed predictive models to evaluate the level of reliability a utility can expect by deploying it, as well as the further improvement needs of the predictive algorithm itself. Accordingly, this paper will review the analyses performed, the outcomes of this study, and discuss plans to improve upon the analyses to ensure that maximum usefulness of the model can be achieved.

1. BACKGROUND

Water demand is increasing rapidly around the world, while the water resources are becoming gradually more scarce. The water network infrastructure is also critically deteriorated, leading to a growing rate of failure of treated water. These failures are not only worsening the imbalance of supply-demand, they also can be accompanied by considerable consequences, such as decreased reliability, supply interruptions, and societal inconveniences (Yerri et al., 2017). Some of these consequences could be prohibitively expensive depending on the size and location of the failed pipe, as well as the impact it can have on the overall function of the system. Considering the critical condition of the aging water infrastructure and the significant number of failures, which is expected to keep increasing in the coming years (Thornton et al., 2008), a more aggressive proactive management approach of water assets is required, rather than the currently more common reactive one.

To prevent catastrophic failures, the goal is to proactively identify and prioritize the high-risk pipes and replace or rehabilitate them in time. However, the success of the prioritization step highly relies on developing a precise failure prediction model which can estimate the probability of failure of a system at the pipe-level. Developing such model can be a problematic task due to the limited availability and quality of the data and significant uncertainty associated with the actual conditions of the buried assets. Several studies have been done recently to tackle this problem; these can be categorized mainly in knowledge-driven physical models and data-driven models (Jenkins et al., 2014; Zhang et al., 2018).

Knowledge-driven models are mostly proposed to predict the deterioration process of the pipeline, focusing mostly on the individual components of the physical process that leads to the failure of the pipe, e.g. corrosion (Rajani and Makar, 2000). However, deterioration of the pipe is usually a complex process, not still fully understandable by the current physical models (Winkler et al., 2018). Additionally, these models can only take into account a few numbers of affecting covariables at the time, which make them applicable only for certain conditions, such as particular pipe materials or failure modes. In data-driven models, on the other hand, the failure pattern is assumed to be the same for pipes that share similar attributes. The failure pattern supposedly can be learnt from the available dataset, such as past pipe failure history of the system and time-dependent and -independent features of the pipe itself. Among data-driven models, artificial intelligence algorithms have recently drawn much attention in the forecasting of water main breaks, owing mainly to the capability of these models to solve complex problems by ingesting a large amount of data without the necessity of detailed model assumptions (Tran et al., 2007; Nicklow et al., 2017; Winkler et al., 2018, Chen et al., 2019). While these models have proven to be beneficial for asset owners, they can only learn from what they see, which means if the model is fed poorly, because of lack of data, lack of cleaned data, or poor-quality data it will only provide poor results. Also, there is usually a necessity of high amount of data and computational resources for these models, which motivates the necessity of reaching a reasonable trade-off among performance and computational burden (Sadrfaridpour et al., 2016).

Accordingly, the predictive model developed in this study aims to estimate the future probability of failure using historical failure patterns, environmental and demographic attributes, as well as pipe characteristics. The goal is to improve the predictive power of the learning algorithms by developing appropriate engineered features from raw data to assist the learning process. In addition to the model improvement, another focus of study is to evaluate the reliability of the model and ensure that the interpreted results are operationally applicable for the asset owners.

2. DATA AND METHODS

Partnered with a utility based in the mid-Atlantic region, the authors have obtained information on pipe break dates and locations between 1999 to 2018, together with the geospatial data of the pipes. The break data has been aggregated to a yearly temporal scale. Environmental and demographic data was also collected from public sources and spatially joined to the pipe network. Historical failure trends were also mined and used as explanatory variables. The variation of some variables over time has been incorporated in the analysis; these have been divided into two groups of “time-independent” and “time-dependent” variables:

- The time independent variables are the fixed covariates over the time windows which have been considered in the analysis. Variables included in this analysis are pipe features

(i.e. geographical locations of each pipe segment, length, diameter, material and installation year), hydraulic characteristics, soil properties (i.e. corrosivity index for concrete and steel, and run-off propensity), land use, and proximity to transportation infrastructures (road and rails);

- The time dependent variables are those that can change value over the course of the observation period. The current predicting model takes into account the effect of the climate data and engineered features. The importance of engineered features and the process of extracting them from raw data will be discussed in more detail in the following subsections.

2.1 Engineered Features

Feature engineering is the process of using domain knowledge of data by extracting features from raw data and transforming them into formats that are suitable for the predictive model (Zheng and Casari, 2018). If the best optimized engineered features are extracted from the dataset, they help facilitate the learning process of the model and increase the predictive power of algorithms.

In case of pipe failure prediction, the analysis is modelled as a binary classification task: failure (= 1) or no failure (= 0). The majority of the pipes in a water network system have zero failures in their records, which makes the precise prediction of a rare event, i.e. having failure, even more challenging. The characteristic of having frequent zero-value observations in a dataset is known as zero inflation. To combat the zero-inflation problem, a temporal lag and spatial buffer have been defined to introduce correlation between observations. It is a decent assumption, as previous studies have shown a strong spatial and temporal correlation between failures of the pipe in a water network system and its surrounding neighbors. Additionally, the spatio-temporal features were implemented since in practice, typical replacement schedules include more than one pipe, generally small neighborhoods or blocks, depending on the utility's priorities and budgetary considerations.

However, to maximize the potential benefit from the failure pattern recognition using temporal/spatial correlation, the most optimized combination of temporal lag and spatial buffer will be selected, based on evaluating several scenarios. It is worth highlighting that the optimized temporal lag and spatial buffer are specific to each network, in terms of system configuration and its temporal/spatial break history. After selecting the best lag and buffer to explore the spatial and temporal patterns, a linear buffer is usually used for determining the buffered/lagged failure history. This means that the linear distance along the pipe is used to spatially identify failures within a specified spatial buffer of a given pipe as opposed to a radial or rectangular area-based search. Linear spatial scanning is preferred for water distribution system data, as neither circular nor rectangular scan windows take the network structure into account (De Oliveira et al., 2010).

2.2 Algorithm Selection

Once all the inputs are collected, it is important to work on pattern recognition by utilizing different kinds of Machine Learning (ML) algorithms. Usually in the learning process of these algorithms, the machine is presented with example inputs and their actual outputs with the goal to produce an inferred function by learning a general rule that maps the inputs to outputs. The training process usually continues until the model achieves the desired level of accuracy on the training data. One type of desired output in learning models is called Regression in which the

outputs are continuous, rather than discrete. In the matter of probability of failure prediction, the problem, as mentioned before, has been framed as a binary classification task of failure (positive response = 1) or no failure (negative response = 0); the model returns a continuous variable bound between 0 and 1 corresponding to the probability of a positive response. Several algorithm structures are used to determine the mathematical relationship between two or more variables with different levels of dependency between them. The output and performance of each model varies based on the interdependency of variables, as well as the specific structure of the model. The algorithms which have been tested for predicting the probability of water main failures in the current analysis are the following:

1. **Logistic Regression (LR):** an extension of linear regression which applies classification algorithm to assign observations to a discrete set of classes.
2. **Support Vector Machine (SVM):** a pattern classification algorithm which is based on a collection of hyperplanes fitted to the dataset, used to best divide observation classes.
3. **Classification Tree (CT):** recursive partitioning of dataset, structured as a decision tree.
4. **Random Forest (RF):** ensemble of classification trees, each trained with bootstrapped samples of original data and with random subset of variables used for node splitting.

A python package, named ScikitLearn, has been utilized to call each of these algorithms. The goal in all these structures is training the model with the historical data to learn the relationships between different covariates. Furthermore, the aim is to also narrow down the possible relationships and correlations between different variables to the most critical ones to avoid overfitting of the predictive model. While an overfitted predictive model may produce observations that closely match the historical data, it may fail to fit additional data or predict future observations in a reliable manner. Additionally, the objective is to get the most accurate probability of failure estimate for the future.

3. DEMONSTRATION

The case study area that supports the water pipeline system information and the break records provides high quality treated water to approximately 600,000 residents and is comprised of over 2,000 kilometers of water mains, 35,000 watermain valves and 9,000 municipal fire hydrants. The failure data from 1999 to 2018 is provided, as well as raw pipe data. Following common predictive model practice, the dataset has been divided into two different subsets:

- the training dataset, which is the sample of data used to fit the model; the model sees and learns from this data;
- the validation dataset, which is the sample of data used to provide an unbiased evaluation of final model fit on the training dataset and is generally what is utilized to evaluate competing models.

Regarding the case study area, the last three years of break data, i.e. 2016, 2017 and 2018, are separated as the validation dataset.

3.1 Preliminary Performance Evaluation

During training of the model a preliminary step was implemented where several trials of random holdouts were implemented to compare the model structures against each other, while tuning of the parameters was also conducted loosely. To aid the selection of the best model among four algorithms, the traditional precision measure in binary classification has been selected for preliminary evaluation of the models, but focusing only on the top 10^{th} percentile of the observations:

$$Precision_{10} = \frac{TP_{10}}{OBS_{10}} \quad (1)$$

The top 10th percentile refers to the ranking achieved by sorting the outputs, highest to lowest, based on the probability of failure (positive response = 1). That way, according to the predictive results, the assets with the highest likelihood of failure are targeted. In the equation above, OBS_{10} is the total number of pipe segments found in the top 10th percentile and TP_{10} is the number of the true positive observation found in the top 10th percentile of the prediction outputs. In other words, $Precision_{10}$ is the proportion of pipes in the top 10th percentile, ranked according to the predicted probability of positive responses, which are indeed positive when compared to the testing dataset. Based on the network configuration and break history of the system, the optimized spatial buffer of 50 meters has been determined for the case study area. Accordingly, the response of a given pipe has been defined as positive (= 1) if the pipe, or its 50 meter surrounding neighboring pipes, will fail.

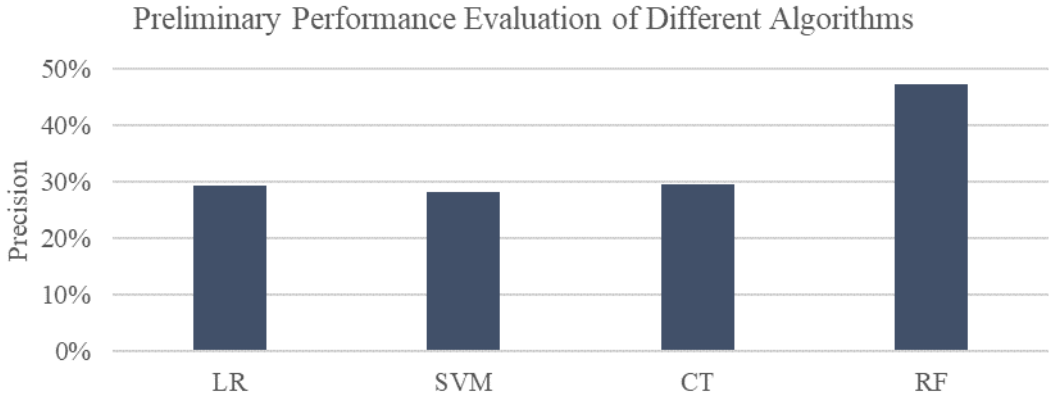

Figure 1- Initial Performance Evaluation of Different Algorithms.

As can be seen from Figure 1, Random Forest (RF) is the most precise model among four algorithms. Accordingly, it has been selected for further analysis and more trials of random holdouts have been performed for tuning the algorithm parameters, e.g. number of trees in the forest and the minimum number of samples allowed in each leaf. Finally, the best RF structure is applied on the entire training dataset and its results are evaluated against the validation data to test the predictive accuracy and precision.

4. PREDICTION RESULTS

The last three years of break data (2016-2018) has been kept separate from the training dataset, as previously mentioned, to evaluate the performance of the developed predictive model. Using the best RF algorithm structure, the outputs are sorted on a pipe-level basis from the highest to lowest probability of failure. Focusing on the top percentiles of the predictions, e.g. 1st or 10th percentiles, the riskiest pipes can be targeted. A top-percentile focus has been utilized since, practically, utilities only have a limited budget for capital improvement. Therefore, the most useful models should be highly accurate and precise for predicting the riskiest assets. Using the binary classification problem, four important terms will be applied to quantify the performance of model over each percentile: True Positives (TP), True Negatives (TN), False