Consideration of pipe service life

Jong Wook Choe
1994wook@gmail.com

Abstract

In this hypothetical scenario, we're given some dummy client data which closely follows what we typically see in the real world. The goal is to clean and do some basic preprocessing on data, as well as provide some insight into how the data are structured.

1 Introduction

"Predict breaks, target leaks and reduce Non-Revenue Water in as little as 45 days."

Rapidly identifying weaknesses in water pipe networks, understanding the impact of an unplanned failure and improving the overall reliability of infrastructure to preserve capital.

Throughout the project, the comparison between variables by the condition of the pipe will be present. In extending the analysis there will be potential modeling ideas that might be effective given what we observed in the raw data.

2 Data Wrangling

The provided pipe file (GM2022_assests.csv) which contains the data describing the actual pipes themselves, and a break file (WO_EXPORT.csv) contains the data describing past breakages that occurred throughout the pipe network.

2.1 Data Description

Pipe File

- native pipe id: ID used by a client to identify pipes. Not necessarily unique.
- asset id: Unique ID set by us. Should be integers 0 to n, where n is the number of rows in the data frame.
- material: Should be on of DIP, CAS, PVC, AC, and SP.
- install yr: Year of installation
- diameter: Diameter of pipe
- abandoned: Indicates whether the pipe is being used or not.
- soil ph: pH level of soil surrounding pipe.
- mean low temp: Annual mean low temp (celsius) in the area.
- soil moisture index: Quantifies water saturation level of soil.

Break File

- native pipe id: ID from pipe file that matches the break to the pipe it occurred on
- asset id: Corresponding unique ID for matching the break to the pipe it occurred on.
- break date: Year of breakage

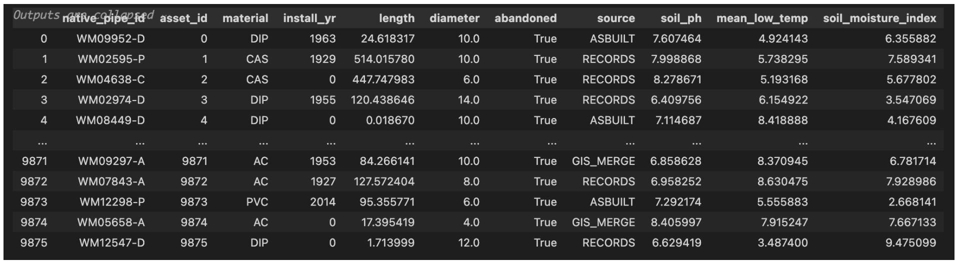
2.2 Data Exploration

Impossible values have been converted to null. All null values then are left as instructed. Also, extra unnecessary columns in the data were also kept since the required ones are present.

Column names exactly follow the convention above to avoid further model interpretation. In addition to that the pipe material has been consolidated into the categories listed in the table.

- DIP stands for ”Ductile Iron”
- CAS stands for ”Cast Iron”
- AC stands for ”Asbestos Concrete”
- PVC stands for ”Polyvinyl Chloride”
- SP stands for ”Steel”

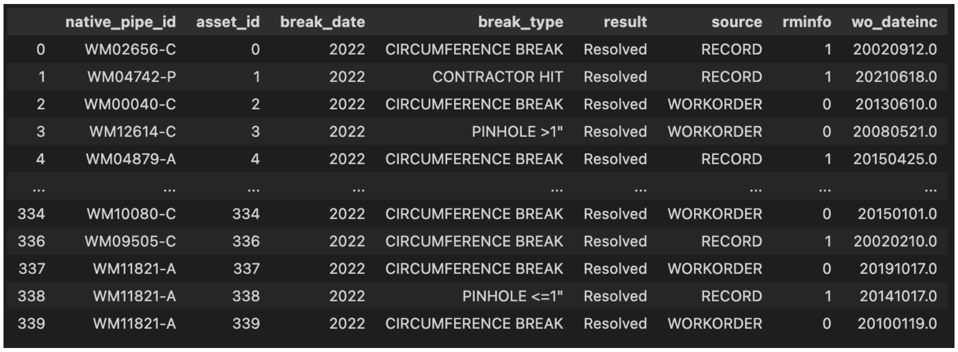
Only the year of installation and breakage is used in the modeling/inference phase. Therefore, the month and date were dropped. Lastly, breaks that aren’t due to natural causes were excluded from the analysis. For instance, a contractor hits when a construction crew accidentally drills into a pipe as it would violate the main hypothesis that the target variable is dependent on the inputs.



Outputs are collapsed

	native_pipe_id	asset_id	material	install_yr	length	diameter	abandoned	source	soil_ph	mean_low_temp	soil_moisture_index
0	WM09952-D	0	DIP	1963	24.618317	10.0	True	ASBUILT	7.607464	4.924143	6.355882
1	WM02595-P	1	CAS	1929	514.015780	10.0	True	RECORDS	7.998868	5.738295	7.589341
2	WM04638-C	2	CAS	0	447.747983	6.0	True	RECORDS	8.278671	5.193168	5.677802
3	WM02974-D	3	DIP	1955	120.438646	14.0	True	RECORDS	6.409756	6.154922	3.547069
4	WM08449-D	4	DIP	0	0.018670	10.0	True	ASBUILT	7.114687	8.418888	4.167609
...	...	...	...	...	...	...	...	...	...	...	...
9871	WM09297-A	9871	AC	1953	84.266141	10.0	True	GIS_MERGE	6.858628	8.370945	6.781714
9872	WM07843-A	9872	AC	1927	127.572404	8.0	True	RECORDS	6.958252	8.630475	7.928986
9873	WM12298-P	9873	PVC	2014	95.355771	6.0	True	ASBUILT	7.292174	5.555883	2.668141
9874	WM05658-A	9874	AC	0	17.395419	4.0	True	GIS_MERGE	8.405997	7.915247	7.667133
9875	WM12547-D	9875	DIP	0	1.713999	12.0	True	RECORDS	6.629419	3.487400	9.475099

Figure 1: Pipe Data frame



	native_pipe_id	asset_id	break_date	break_type	result	source	rminfo	wo_dateinc
0	WM02656-C	0	2022	CIRCUMFERENCE BREAK	Resolved	RECORD	1	20020912.0
1	WM04742-P	1	2022	CONTRACTOR HIT	Resolved	RECORD	1	20210618.0
2	WM00040-C	2	2022	CIRCUMFERENCE BREAK	Resolved	WORKORDER	0	20130610.0
3	WM12614-C	3	2022	PINHOLE >1"	Resolved	WORKORDER	0	20080521.0
4	WM04879-A	4	2022	CIRCUMFERENCE BREAK	Resolved	RECORD	1	20150425.0
...	...	...	...	...	...	...	...	...
334	WM10080-C	334	2022	CIRCUMFERENCE BREAK	Resolved	WORKORDER	0	20150101.0
336	WM09505-C	336	2022	CIRCUMFERENCE BREAK	Resolved	RECORD	1	20020210.0
337	WM11821-A	337	2022	CIRCUMFERENCE BREAK	Resolved	WORKORDER	0	20191017.0
338	WM11821-A	338	2022	PINHOLE <=1"	Resolved	RECORD	1	20141017.0
339	WM11821-A	339	2022	CIRCUMFERENCE BREAK	Resolved	WORKORDER	0	20100119.0

Figure 2: Break Data frame

2.3 Data Visualization

Data visualization is a key skill of any data expert and applied in a wide variety of areas from scientific research to industrial application.

The following bar plot shows the count of pipe installments by year. Based on data the pipes were mostly built from 1930 to 1960. Also, there was a large gap between 1970 to 2010.

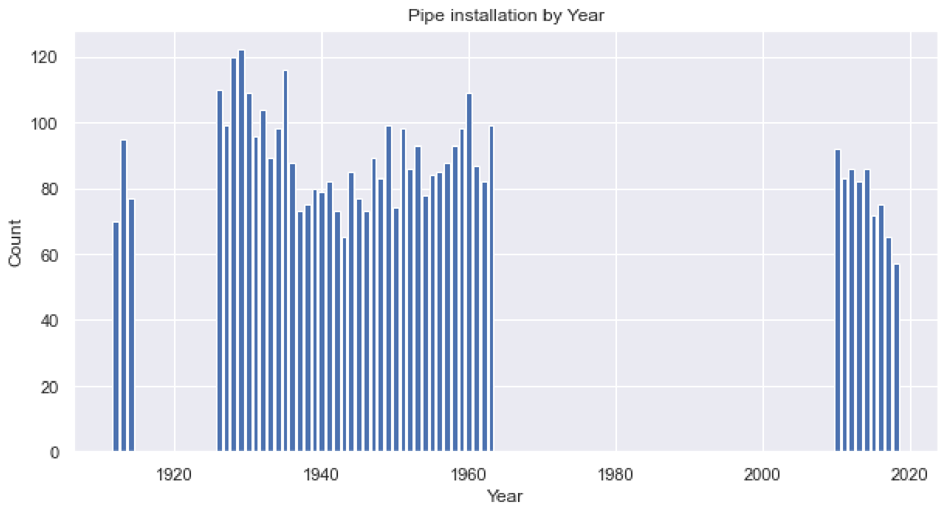


Figure 3: Pipe count barplot by Year

Moving on following five box plots represent distributions of various variables in the data that differ by pipe condition. Bar charts are appropriate for counts, whereas box plots should be used to represent the characteristics of a distribution. The left side (0) represents the pipes that were not noticed to be broken. The other side (1) represents the pipes that were reported to be broken.

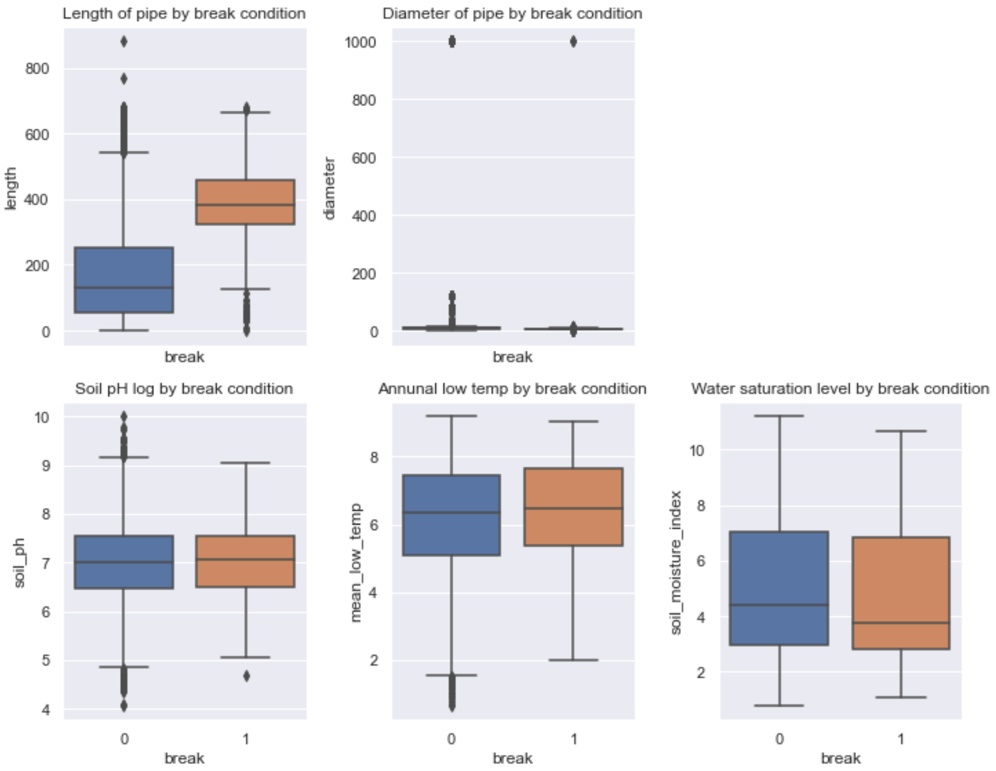


Figure 4: Boxplot Distribution of variables

Based on the box plot a lot of features seem to have similar distribution between broken or unbroken pipes. Some notable features are length, their means are pretty different. The mean of unbroken pipes was below 200 while broken pipes were close to 400. It could easily be interpreted as longer pipes are more hazardous than shorter pipes, however, there are large outliers from unbroken pipes that have extensive length. Therefore, it is more appropriate to translate that middle-size pipes are the riskiest.

Going next following plots show the relationship between a numerical and one or more categorical variables using visual representations.

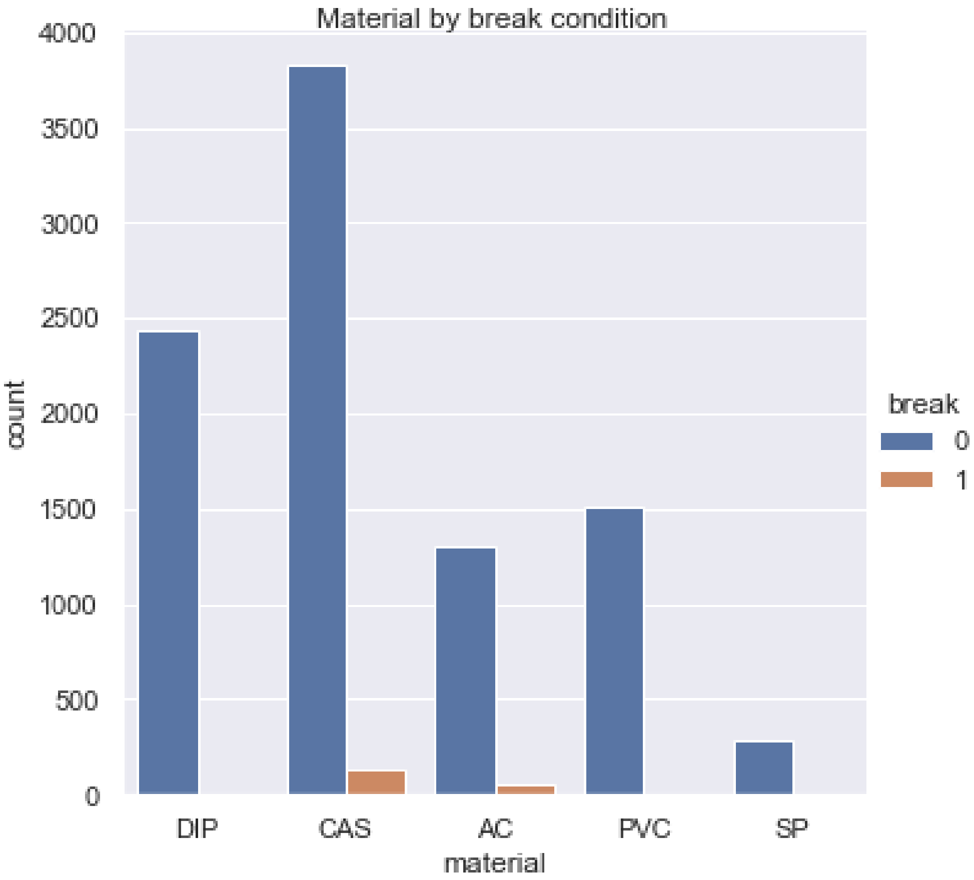


Figure 5: Material of pipe Barplot by break condition

To begin with, the above plot represents the material of pipes by break condition. Pipes were made of ductile iron, cast iron, asbestos concrete, polyvinyl chloride, and steel. In large part, pipes were made with cast iron and ductile iron going behind. An interesting part is asbestos concrete was not largely used but there were noticeable counts of broken pipes.

To see more clearly the following bar plot represents the proportion of broken pipes. The percentage of each material represents the ratio of broken pipes over total pipes.

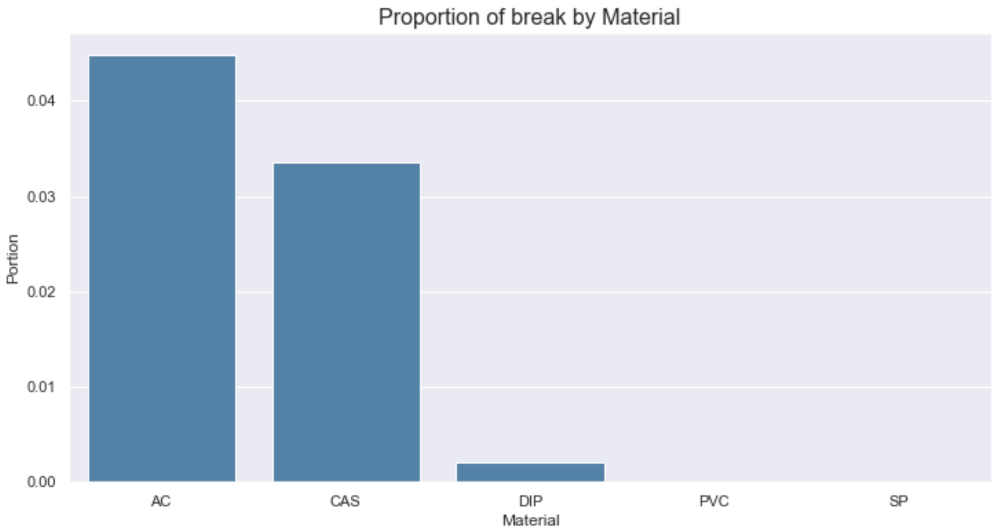


Figure 6: Percentage Barplot of break pipe

As mentioned earlier, asbestos concrete showed significance among other materials. It is significant since it is not the most frequently used material. Therefore, keeping an eye on asbestos concrete pipes will be an option. Another notable part is cast iron which is the most frequently used material. Although the percentage is very low it seems using ductile iron is more beneficial in the future.

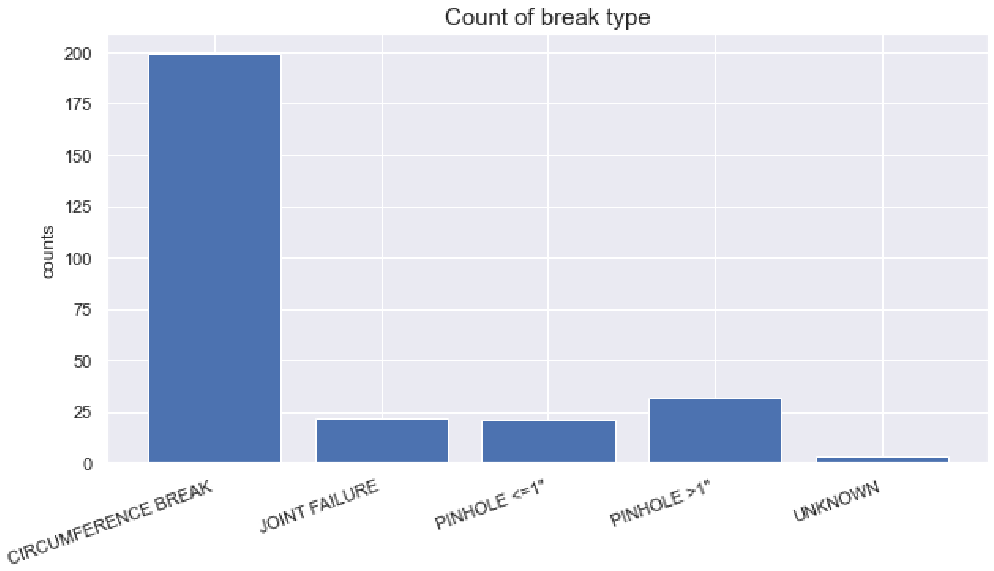


Figure 7: Break type Barplot

Lastly, the above bar plot shows the counts of different break reasons. Only with this plot, we can't provide the exact characteristics of pipes that have broken. However, a notable feature is that joint failure or pinhole size was not the most affecting factor.

3 EDA

3.1 Correlation

Negative correlations between two variables mean as one variable increases, the other variable decreases. On the opposite positive correlations mean as one variable increases, the other one increases as well. The following heat map is a graphical representation of data that uses a system of color coding to represent correlations of variables.

The target variable is the ”break” column which is a categorical variable in the pipe statistics table of the pipe condition. The target variable was made based on the break statistics table to 1 and 0. Explanatory variables were selected from the pipe table.

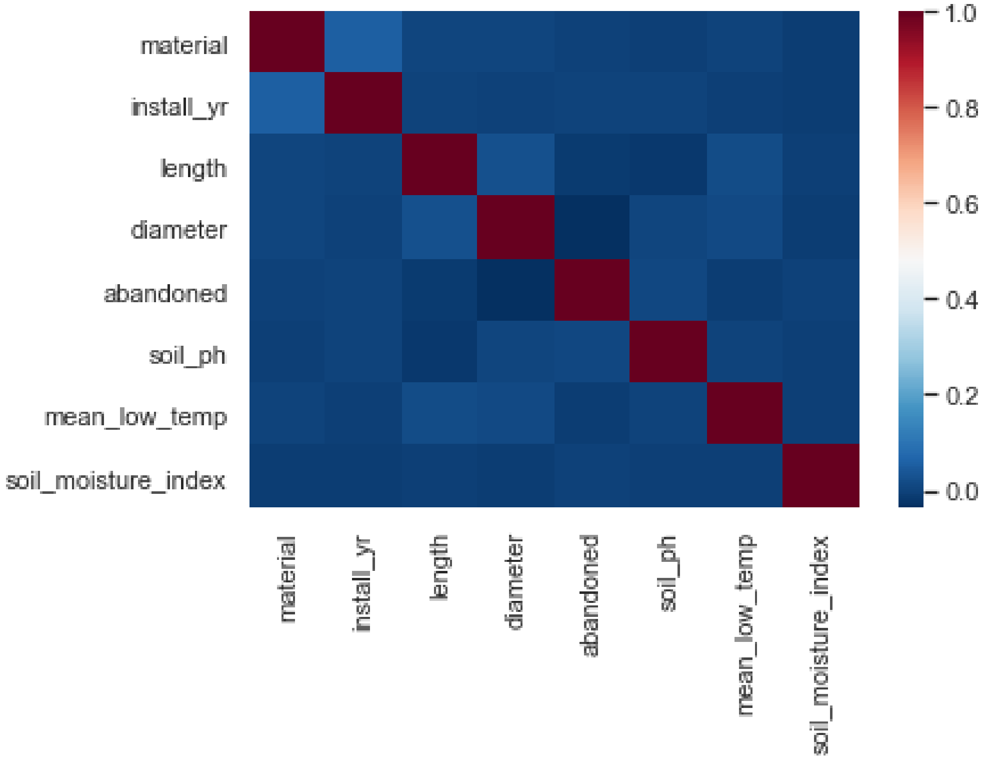


Figure 8: Correlation Heatmap

The above graph shows the correlation between each explanatory variable in a heat map. If the coefficient value lies between ±0.50 and ±1, then it is said to be a strong correlation. When the value lies below ±0.29, then it is said to be a small correlation. A weak correlation indicates that there is a minimal relationship between the variables.

3.2 Coefficient

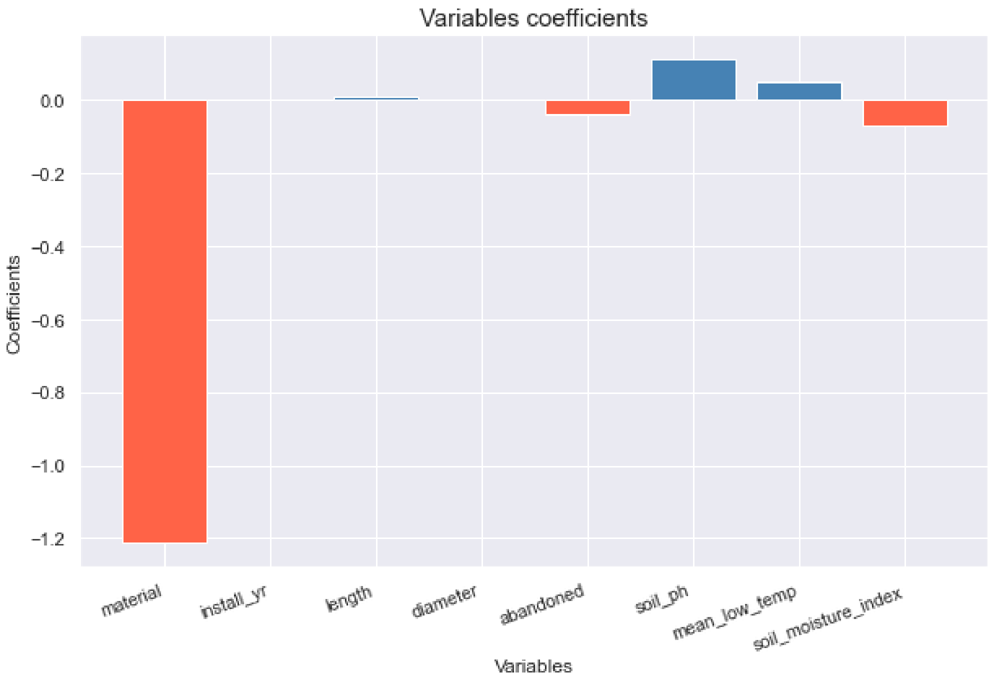


Figure 9: Feature Coefficients

The bar plot above shows the coefficients of each variable. The categorical variable ”Material” is converted to numerical variables.

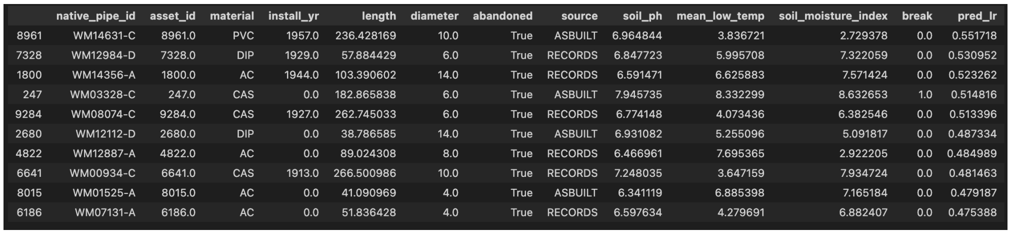
$$Material = \begin{cases} 0, & \text{if } Material = AC \\ 1, & \text{if } Material = CAS \\ 2, & \text{if } Material = DIP \\ 3, & \text{otherwise} \end{cases} \tag{1}$$

Based on that material variable had the highest coefficient. Therefore, a pipe made of AC(Asbestos Concrete) stat has a higher probability to be broken.

4 Data Modeling

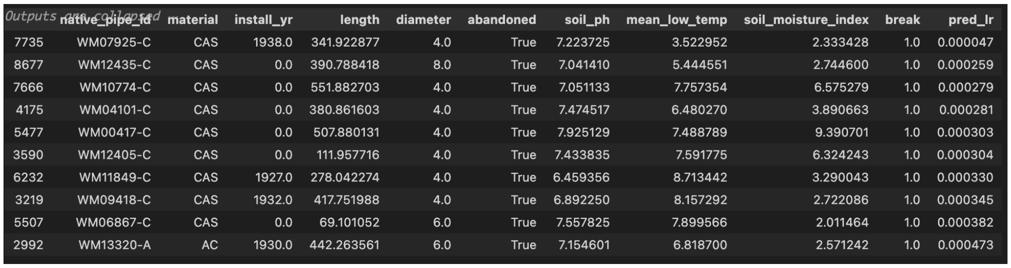
Since our target variable is a binary categorical variable, Logistic regression to train the model would be a good option. With this modeling, we were able to get the probability of the pipes that are possibly pipe breaking.

After training the Logistic regression model, the following table shows the Top 10 and lowest 10 pipes that are possible to break.



	native_pipe_id	asset_id	material	install_yr	length	diameter	abandoned	source	soil_ph	mean_low_temp	soil_moisture_index	break	pred_lr
8961	WM14631-C	8961.0	PVC	1957.0	236.428169	10.0	True	ASBUILT	6.964844	3.836721	2.729378	0.0	0.551718
7328	WM12984-D	7328.0	DIP	1929.0	57.884429	6.0	True	RECORDS	6.847723	5.995708	7.322059	0.0	0.530952
1800	WM14356-A	1800.0	AC	1944.0	103.390602	14.0	True	RECORDS	6.591471	6.625883	7.571424	0.0	0.523262
247	WM03328-C	247.0	CAS	0.0	182.865838	6.0	True	ASBUILT	7.945735	8.332299	8.632653	1.0	0.514816
9284	WM08074-C	9284.0	CAS	1927.0	262.745033	6.0	True	RECORDS	6.774148	4.073436	6.382546	0.0	0.513396
2680	WM12112-D	2680.0	DIP	0.0	38.786585	14.0	True	ASBUILT	6.931082	5.255096	5.091817	0.0	0.487334
4822	WM12887-A	4822.0	AC	0.0	89.024308	8.0	True	RECORDS	6.466961	7.695365	2.922205	0.0	0.484989
6641	WM00934-C	6641.0	CAS	1913.0	266.500986	10.0	True	RECORDS	7.248035	3.647159	7.934724	0.0	0.481463
8015	WM01525-A	8015.0	AC	0.0	41.090969	4.0	True	ASBUILT	6.341119	6.885398	7.165184	0.0	0.479187
6186	WM07131-A	6186.0	AC	0.0	51.836428	4.0	True	RECORDS	6.597634	4.279691	6.882407	0.0	0.475388

Figure 10: Top 10 probability of pipe break



Outputs are collapsed

	native_pipe_id	material	install_yr	length	diameter	abandoned	soil_ph	mean_low_temp	soil_moisture_index	break	pred_lr
7735	WM07925-C	CAS	1938.0	341.922877	4.0	True	7.223725	3.522952	2.333428	1.0	0.000047
8677	WM12435-C	CAS	0.0	390.788418	8.0	True	7.041410	5.444551	2.744600	1.0	0.000259
7666	WM10774-C	CAS	0.0	551.882703	4.0	True	7.051133	7.757354	6.575279	1.0	0.000279
4175	WM04101-C	CAS	0.0	380.861603	4.0	True	7.474517	6.480270	3.890663	1.0	0.000281
5477	WM00417-C	CAS	0.0	507.880131	4.0	True	7.925129	7.488789	9.390701	1.0	0.000303
3590	WM12405-C	CAS	0.0	111.957716	4.0	True	7.433835	7.591775	6.324243	1.0	0.000304
6232	WM11849-C	CAS	1927.0	278.042274	4.0	True	6.459356	8.713442	3.290043	1.0	0.000330
3219	WM09418-C	CAS	1932.0	417.751988	4.0	True	6.892250	8.157292	2.722086	1.0	0.000345
5507	WM06867-C	CAS	0.0	69.101052	6.0	True	7.557825	7.899566	2.011464	1.0	0.000382
2992	WM13320-A	AC	1930.0	442.263561	6.0	True	7.154601	6.818700	2.571242	1.0	0.000473

Figure 11: Lowest 10 probability of pipe break

However the prediction rate of the machine learning model is too low, there are several things you can try to improve it:

Feature Engineering: It is possible that the features you are using are not informative enough to predict the target variable. You can try to engineer new features or select better features to improve your model's performance.

Model Selection: Try different models to see if they perform better on your data. Sometimes, a different model can give much better results than the current one.

Hyperparameter Tuning: Every machine learning model has some hyperparameters that can be tuned to improve its performance. Try different hyperparameters to see if they improve your model's performance.

Data Preprocessing: Your data might have some outliers, missing values, or other issues that are affecting your model's performance. You can try different data preprocessing techniques to see if they improve your model's performance.

Increase Data Size: Sometimes, a model might not have enough data to learn from. In this case, you can try to increase the amount of data to see if it improves the model's performance.

5 Conclusion

Throughout the analysis, there were some characteristics of pipes that have broken. Based on the box plot a lot of features seem to have similar distribution between broken or unbroken pipes but the length of the pipes showed the difference. Also found out the material of pipes could affect the pipes break. Pipes were mostly made of cast iron, however, the pipes made of asbestos concrete had the highest proportion of break pipes.

Moreover, using the Logistic regression model possible pipes of failure have been found. Setting the pipe condition as the target variable and setting others as the explanatory variable of the data were trained. However, the prediction rate is low therefore using other models such as Decision tree and Random forest could be a better option for future analysis.