Predicting Utility Pipe Structural Scores

Elyse Cheung-Sutton

Abstract

The purpose of this study is to test whether pipe attributes (System Type, Diameter, Length, Age, Material Type, Operational Score) can be used to predict pipe defects (Structural Score). Current utility pipe data from the City of Sacramento is used in a Decision Tree Classifier. Although the trained model has a 72% accuracy rating, it is unable to accurately predict high Structural Scores and thus cannot be used to predict pipe defects.

Motivation

Cities maintain intricate networks of pipes which convey wastewater to treatment plants and drainage to open waters. This infrastructure is critical to maintaining public health and safety and failures can be catastrophic in terms of construction costs, property damage, and impacts to daily operations. Defects, such as cracks, holes, and offsets increase the likelihood of pipe failure and must be identified in a timely manner.

Currently, maintenance workers log a Structural Score for each pipe based on the severity of defects seen in CCTV video recordings. This process is time consuming and it is nearly impossible to get full coverage in a network with over 100,000 pipes. Using machine learning to predict pipe defects could help fill in null values, reduce maintenance crew workload, and help prioritize maintenance on pipes that are classified as more likely to fail.

Dataset

The dataset is pipe data for the city of Sacramento which includes wastewater, drainage, and combined system pipes. 100,375 records and 8 columns (Facility ID, System Type, Diameter, Length, Age, Material Type, Operational Score, Structural Score) are included in the initial dataset.

Data Definitions:

1. System Type: type of water being conveyed
    - Wastewater and drainage pipes only convey their respective water types; combined pipes convey both wastewater and drainage together.
2. Operational Score: operability of a pipe, PACP* standard scores 0-5
    - High scores indicate reduced operability due to roots, blockages, etc.
3. Structural Score: structural integrity of a pipe, PACP* standard scores 0-5
    - High scores indicate reduced structural integrity due to defects such as cracks, holes, misalignments etc.

*PACP = Pipeline Assessment Certification Program, North American standard for pipe defect identification

Data Preparation and Cleaning

- Records without a Structural Score (predicted variable) were dropped, reducing the record count from 100,375 to 20,365.
- Nulls in ShapeLength column (length of pipe in GIS) were filled using the “SurveyedFootage” column (length of pipe from field observation).
- Nulls in Diameter and Age column were filled using the median value from that system type (i.e. missing diameter values for sewer pipes were set to 6” which was the median value for diameter of all sewer pipes).
    - Median was chosen over mean because outliers skew mean values.
- System Type was merged from 5 groups to 3 groups. “Combined Sewer” and “Combined Drainage” pipes were labeled “Sewer” and “Drainage” respectively.
    - These pipes are the same functionally but are labeled differently due to state mandate.
- Uncommon Material Types (count<50) were labeled as “Other” to reduce encoding in analysis.

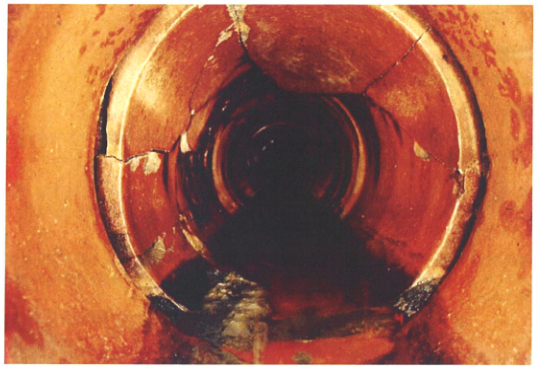
Research Question(s)

Can pipe attributes, such as age, diameter, length, material, operational score, and system type, be used to predict pipe defects (high structural score)?

Examples of Pipes with High Structural Scores

Broken Void Visible (BVV)
PACP Grade 5

Broken Soil Visible (BSV)
PACP Grade 5



Joint Angular Large (JAL)
PACP Grade 4

Methods

- Classification with the Decision Tree Classifier method was used because the predicted variable is categorical and has 6 possible, defined labels (structural scores of 0 through 5, with 0 being no defects and 5 being many defects).
    - 2/3 of the data was used to train the model and 1/3 was used the test the model.
    - Max leaf nodes was set to 10 following the class example.
- Categorical variables (System Type and Material) were converted to numerical values using One Hot Encoding and Feature Hashing. Both methods were chosen for programming simplicity and resulted in the same accuracy score.

Dataframe after One Hot Encoding

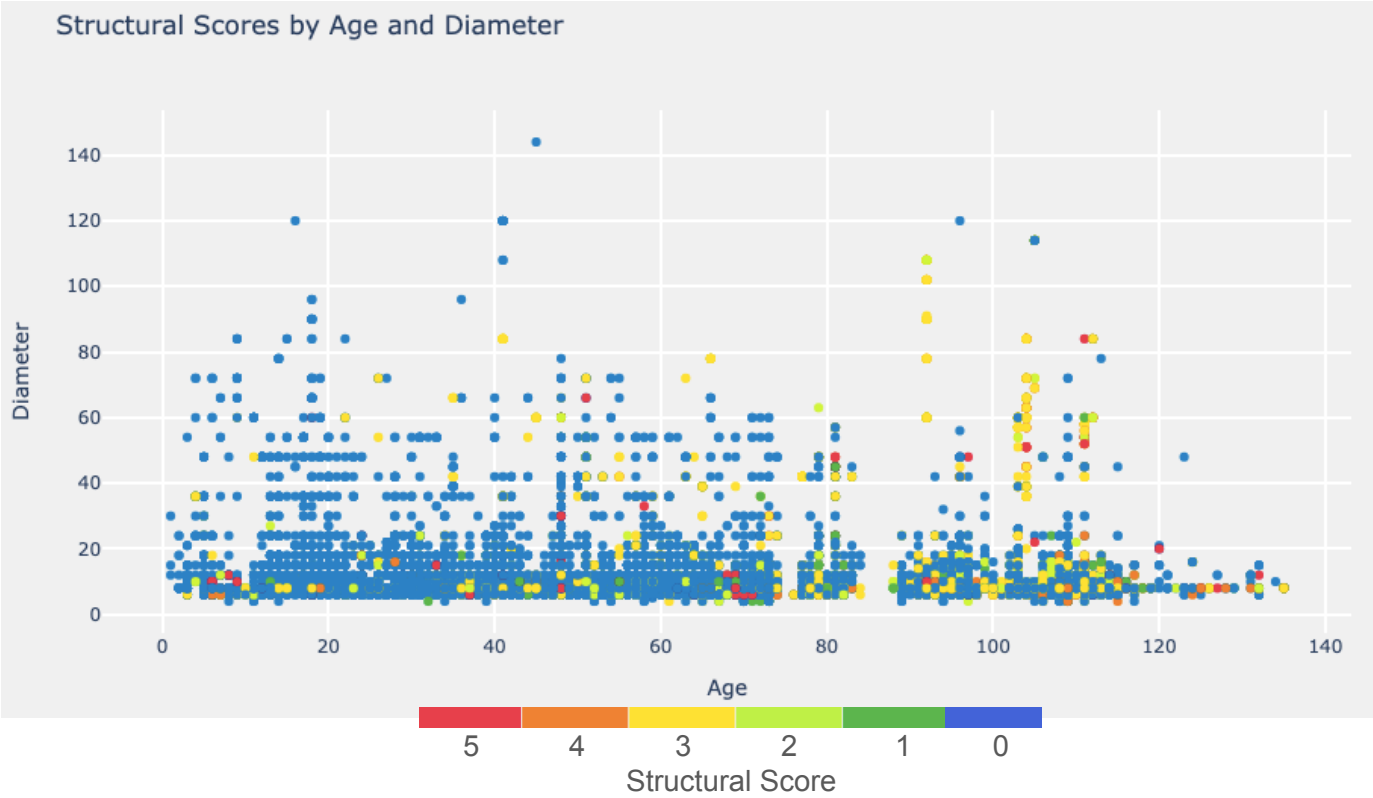
	DIAMETER	SHAPELENGTH	AGE	MAXOPSMAINTSCORE	C	D	S	AC	CMP	CP	CT	Other	PVC	RCP	SP	UNK	VCP
16304	24	16	6	0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
79	12	322	79	0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
12119	12	260	54	0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
14147	12	159	28	0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
5640	6	269	29	0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

Findings

The correlation coefficients for the numeric attributes show that there is a slight positive correlation between Age, Operational Score, and Structural Score. This relationship makes sense because older pipes may be more corroded and deformed leading to structural issues.

	Diameter	Length	Age	Operational Score	Structural Score
Diameter	1				
Length	0.1	1			
Age	-0.02	0.03	1		
Operational Score	0.07	0.19	0.25	1	
Structural Score	0.08	0.15	0.36	0.25	1

Findings

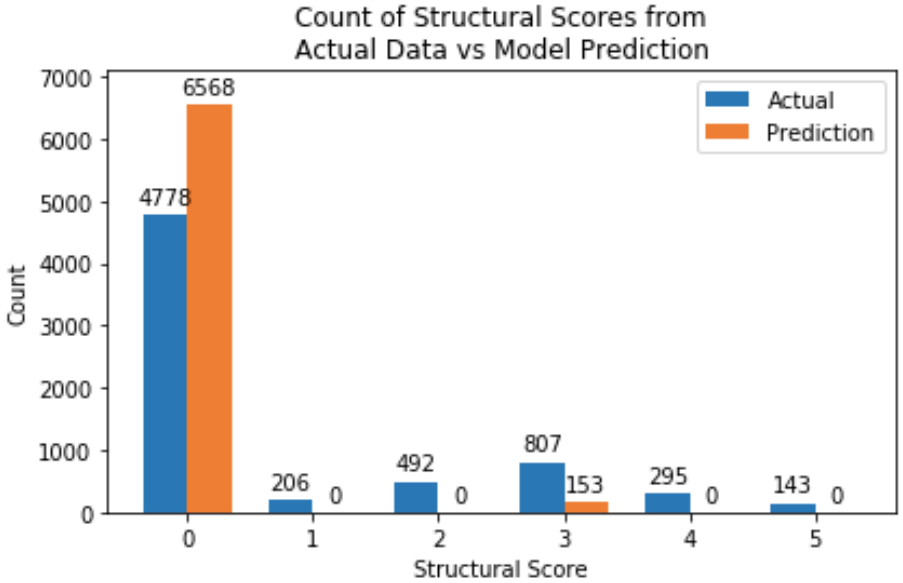


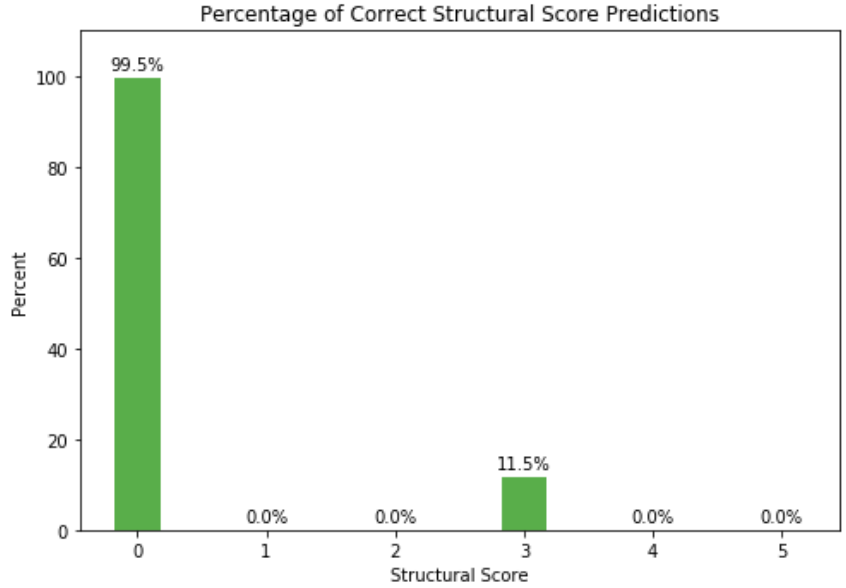
This chart further displays the correlation between high Age and high Structural Score (red, orange, yellow dots*). The correlation is not strong however as red dots (Structural Score = 5) are dispersed throughout.

*I couldn't figure out how to get the Structural Score legend embedded properly in this chart.

Findings

Using the Decision Tree Classifier to predict Structural Scores resulted in a model with 72% accuracy. Upon further investigation, however, the accuracy score is misleading because the model only predicted 0 and 3 values, as shown in the left chart. While the model correctly predicted 99.5% of the 0 scores, it was completely inaccurate in predicting the full range of scores.





Limitations

This project was limited by the input dataset because it only included pipes in the current system. As demonstrated by the high occurrence of “0” Structural Scores (pipes without defect) in slide 9, most of the pipes in the streets are functioning and do not have defects. Historical data containing information for pipes that failed and were replaced may have improved the training data for the model and the model’s predictive accuracy.

This project was also restricted by my limited knowledge of machine learning algorithms. There may have been settings in the Decision Tree Classifier that could have improved model accuracy. There may also be other algorithms that are more appropriate for this particular question and dataset.

Conclusions

The Decision Tree Classifier was not able to predict pipe defects (high structural scores) using the input pipe records and attributes. While the model predicted 0 Scores well, it failed at predicting the more important higher scores. Assessing pipe condition and managing replacement schedules require an accounting of pipe defects, thus, the Structural Scores should still be generated manually through visual inspection.

Hopefully, I will learn of other machine learning algorithms that can better analyze this dataset in future machine learning classes!

Acknowledgements

I acquired the dataset through my current position at the City of Sacramento. I used my experience in this position to understand the data and formulate the research question.

I received feedback on my presentation from Neely B and Jonathan O.

References

NASSCO Pipeline Assessment:
https://www.nassco.org/content/pipeline-assessment-pacp