Title: Water Main Breaks in Madison
Name: ___________

The Madison Water Utility delivers water to the city via a network of water main pipes. Sometimes these mains break and must be repaired. Our goal in this project is to understand seasonal impacts on breaks and predict when a broken pipe is likely to break again in the near future. These predictions have implications for replacement and preventative maintenance.

The City of Madison has published a 23-year history of 7002 breaks online at https://data-cityofmadison.opendata.arcgis.com/datasets/water-main-breaks. The dataset has 46 columns describing each break, including pipe depth, radius, surrounding soil type, and the location of the break. For many records, a number of fields are not populated (for example, the date of the break is missing for 46% of the records).

We select 5 features for our analysis: (1) pipe depth, (2) pipe size, (3) soil type, (4) number of prior breaks, and (5) whether a break occurred in winter. Pipe depth and size were often missing; we imputed the missing values with the most common values for these fields. Soil type was unstructured, with entries such as "Clay with boulders" and "mud/dirt." We translated these descriptions to four columns, clay, rock, sand, and gravel, each containing a 1 when the term appeared. For each break record, we computed the number of prior breaks appearing earlier in the dataset. We set "winter" to 1 when the month is Dec, Jan, or Feb (and 0 otherwise).

We also add a "breaks again" column indicating whether there is another break in the next 5 years following a break. As this time frame extends into the future for the most recent rows, we drop the last 5 years of records from our dataset (it remains to be seen whether these pipes will break again). We also drop rows for which the date is unknown. Sometimes a single pipe has multiple breaks for the same day; in this case, we only keep the last row (dropped same-day rows still contribute to prior break counts, though). These steps reduce our dataset from 7002 to 3361 rows.

We start by exploring the impact of winter on breaks. The blue line of **Figure 1** shows the average number of breaks per month, which range from 51 (April) to 755 (January). 66% of all breaks occur in winter (28% in January alone). The orange line of the figure shows the subset of breaks (31% of them) that will be followed by another within the next 5 years. The pipes that will break again have a similar seasonal pattern to the overall trend: 69% are in winter, and 28% in January, specifically.

We next explore the dimensionality of our data by performing a principal component analysis over our 8 feature columns (although we have 5 features, soil type is represented as 4 columns). **Figure 2** (blue line) shows the result: 2 columns capture 76% of the variance. However, this is largely because the pipe size and depth columns have a larger magnitude than our 0-or-1 columns. After performing standard scaling (orange line), we observe that it is not possible to reduce the dimensionality of our data without losing significant information.

We want to forecast whether a broken pipe will break again in the next 5 years. We create an sklearn pipeline consisting of (1) a SimpleImputer (using the "most frequent" strategy) to fill in missing pipe dimensions, (2) a StandardScaler, and (3) a LogisticRegression. We perform a 75/25% train/test split, stratifying on the "breaks again" column. The model achieves 71% accuracy (not impressive considering 69% of broken pipes don't break again in the time frame). Recall is a 11% (the model rarely predicts a pipe will break again), and precision is 58%.

For most applications we envision (e.g, identifying mains for preventative maintenance), false positives are more acceptable than false negatives, so we re-train our model, this time using the "balanced" class strategy for the LogisticRegression to give more weight to pipes that break again. Although this reduces accuracy to 65% and precision to 43%, the new model is able to identify a majority of pipes that will break again in the future (55% recall), suggesting it would be useful for prioritizing replacement and maintenance.

Figure 3 shows the coefficient weights for each of the features used by the model. We observe that the most important factor for predicting future breaks is how many times a pipe has broken in the past: the more times it has broken before, the more likely it is to break again within the next 5 years. We also observe that winter is a relatively unimportant factor and that pipes laid in sand that break are less likely to break again in the future.

To conclude, our analysis shows that two thirds of all breaks occur during one fourth of the year, the winter months of December, January, and February. However, these pipes that break in winter are not much more vulnerable to future breaks (in the next 5 years) than pipes that break at other times of the year. Rather, the number of previous breaks is the best indicator for whether a pipe will break again soon; we recommend using this count to prioritize maintenance and replacement.

Figures

Figure 1: Seasonal Effect on Breaks

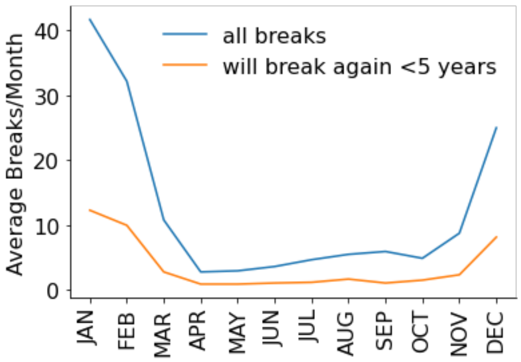


Figure 2: Principal Components of Breaks

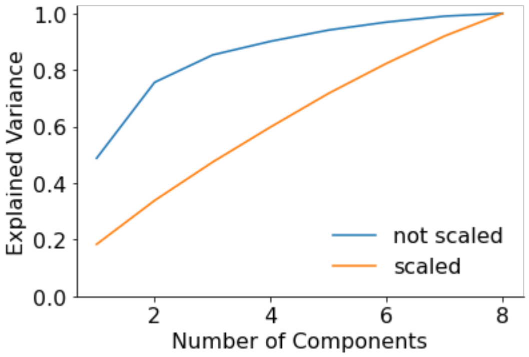
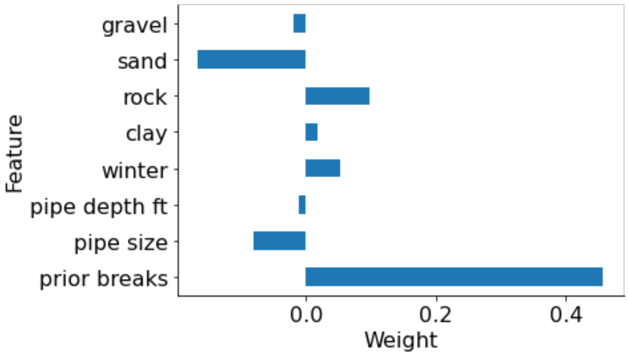


Figure 3: Logistic Regression Coefficients


Feature
gravel
sand
rock
clay
winter
pipe depth ft
pipe size
prior breaks
Weight