Capstone Proposal

Daniel Siegel
May 30th, 2018

Proposal

Domain Background

The Water Point Data Exchange (WPDx) is a global platform for sharing data on rural water distribution points in the developing world. It has a webform that Government agencies can use to upload their records (typically spreadsheets) and connects directly to the monitoring platforms like mWater and Akvo that non-governmental organizations use. The WPDx provides maps, and an API that can be used to query the shared database. This is a valuable resource for underfunded water ministries, which typically do not have the expertise to develop their own data sharing platform, and for the non-governmental organizations, which can now see their projects in the context of what everyone else is doing. It is also a unique dataset that is being used to develop analytic tools and guide resource allocation the water sector.

A team of students at Berkeley did some analysis on this dataset as part of a Data for Good grant. However, their project limited itself to examining each country one-by-one. By treating the whole WPDx as a single dataset, it might be possible to learn more interesting things about the nature of water point durability.

Problem Statement

Sending staff out into the field to inspect water distribution points is time-consuming and expensive. If an algorithm could predict which water points are most at risk of failure, then it would be possible to allocate maintenance and repairs more efficiently. This algorithm would classify a water distribution point as “broken” or “functional” based on attributes about its location, age, and construction.

Datasets and Inputs

The water infrastructure points contained in the WPDx will be used as training data. As an initial proof-of-concept, only the top 7 contributing countries will be used. This represents 350,000 points in total. The relevant features that the WPDx collects on each record include:

- Water Source Type (water_source)
- Water Point Technology (water_tech)
- Installer (installer, source)
- Manager (management, pay)
- Age (install_year, age_years)
- Location (lat_deg, long_deg, country_name, adm1)
- Status (status_id, report_date)

Only one of these attributes (age_year) is not part of the original dataset. It will be calculated by subtracting install_year from the year that status_id was recorded (report_date).

Solution Statement

I will develop a classification algorithm that predicts whether a water distribution point is broken. In other words, the goal is to predict the status_id based on all of the other features.

Benchmark Model

By calculating an average age at which water points break down, I can build an algorithm that classifies water points as broken if they are older than that age, otherwise assuming they are functional. This is essentially a decision tree that only considers one attribute (age_years). By testing how accurately this algorithm can predict water point status, I will have a useful benchmark for comparing more advanced algorithms too. If these algorithms cannot outperform the benchmark model, it means that the attributes in the WPDx are useless for this task, and simply repairing water points when they reach a given age is best we can do.

Evaluation Metrics

The consequence of a false positive is that we will believe a point to be broken, and send a repair crew in error. This is a waste of money and resources, but still better than the consequence of a false negative. Classifying water points as functional when it is actually broken means a repair crew won’t be sent, and the people living nearby will have no access to clean drinking water. With this in mind, I will use an F-score that has β=2 as my evaluation metric. This favors models with high recall, minimizing false negatives without ignoring the importance of precision entirely.

Project Design

The first step is to calculate the age of each water point at the time its status was last evaluated and add this feature to the dataset. Then the numeric features will be normalized and the categorical features like water_tech and management will be vectorized.

Next step is to work with each country individually. They will be separated into training, cross-validation, and testing datasets. I will test out many different classification algorithms, including decision trees, random forests, and logistic regression, and whichever one has the highest F-score I will use on the full dataset.

The key questions I want to examine after training on the full dataset are

- Does training on the full dataset increase accuracy? Or is it better to build a separate model for each country?
- Does the algorithm predict status more accurately than the benchmark model?