Predicting Water Pump Operability in Tanzania

Claude Fried

October 2020

Problem Statement

How can we use information about these pumps to predict if they are working or not?

- The goal is to use data on water pumps in Tanzania to predict the condition of the well.
- There are thousands of water pumps in the country of Tanzania, many of which are an integral part of their area's sustainability.
- Many of these pumps are known to need maintenance (or are not working entirely).

Goal / Value

Goal:

- Our goal is to build a model that can accurately predict the condition of a water pump in Tanzania.

Value:

- Make data-driven decisions on how **to allocate resources** in order to maintain / fix pumps that are most in need.
  - **Save money** – spend only on pumps which are most in need of repairs.
  - **Save time** – avoid erroneously targeting pumps which are functional.

- **Improve maintenance operations** for these water pumps and significantly improve many people's quality of life.

Methodology – OSEMN

Obtain – Scrub – Explore – Model – iNterpret

1. Obtain: Gather data.
2. Scrub: Clean data.
3. Explore: Get to know the data. Engineer features.
4. Model: Train, test, validate, and compare models.
5. Interpret: Glean insights from the model's predictions.

Methodology – OSEMN

Obtain

- The data has been collected from Taarifa (*http://taarifa.org/*) and the Tanzanian Ministry of Water.
- 59,000 training data points.
- 41 original features.

Methodology – OSEMN

Scrub

Examples:

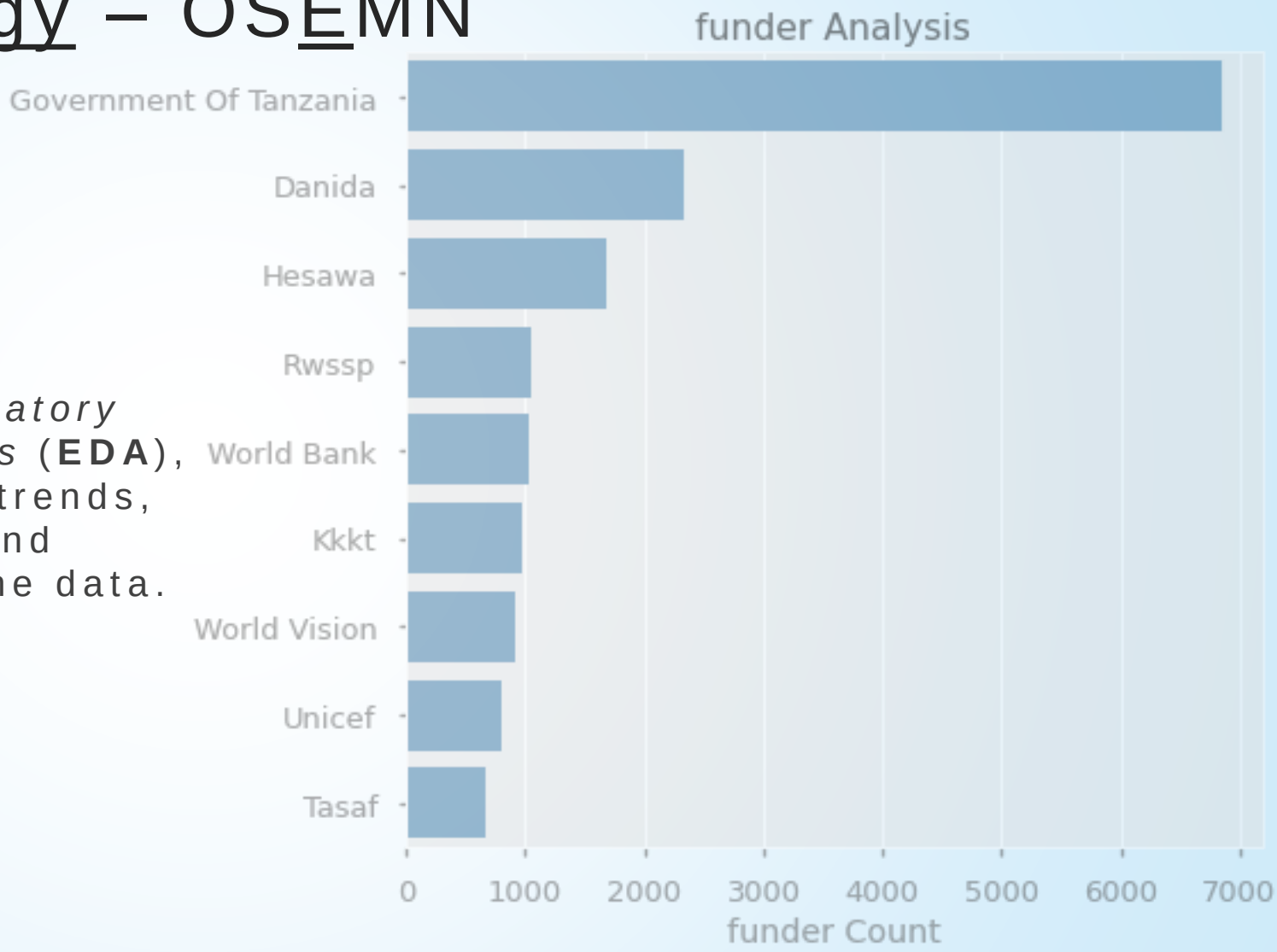
- Format each feature into the correct data-type.
  - e.g.: Is this feature numerical or categorical?

- Determine what to do with…
  1. Missing values for each feature.
  2. Numerous categories within a feature.
  3. Duplicate / similar features.

Methodology – OSEMN

Explore

- Using *exploratory data analysis* (**EDA**), we can find trends, groupings, and outliers in the data.



Methodology – OSEMN

Model

The modeling process included several steps.

1. Selecting which model-types to train.
2. Searching for optimal settings for these models.
3. Creating a new model made up of these trained and optimized models.
4. *Select a final model from which we will gain insights.*

Methodology – OSEMN

Final Model – Random Forest

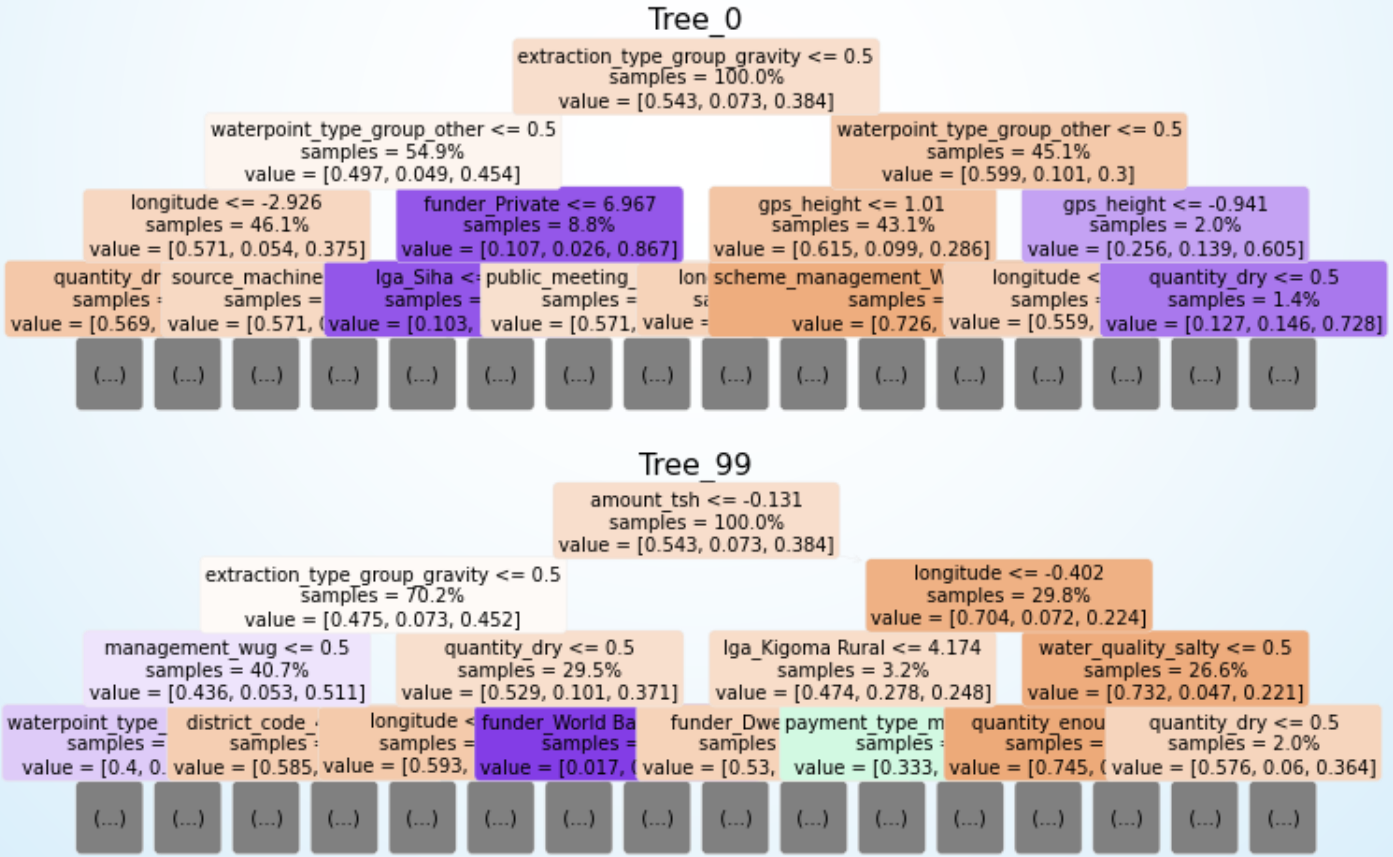
The model that best-balanced performance and speed was the **Random Forest Classifier.**

Random Forests use a collection of *Decision Trees* to make predictions.

Methodology – OSEMN

Decision Trees

A *Decision Tree* takes all the data and splits it numerous times into smaller and smaller groups until each group contains only a single data point.

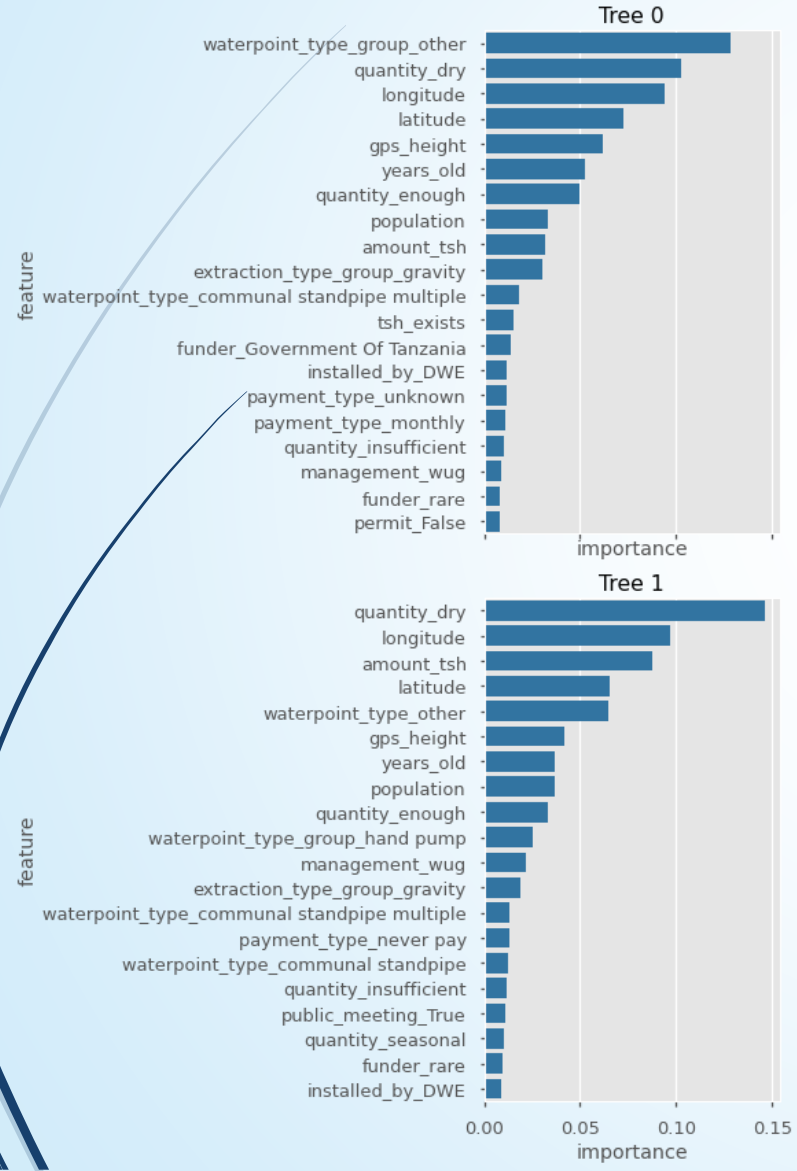


Here are the first three splits of two unique Decision Trees.

The Forest that was used to make final predictions and insights uses 100 unique trees which split the data (over 44,000 points) until each split contains a single entry.

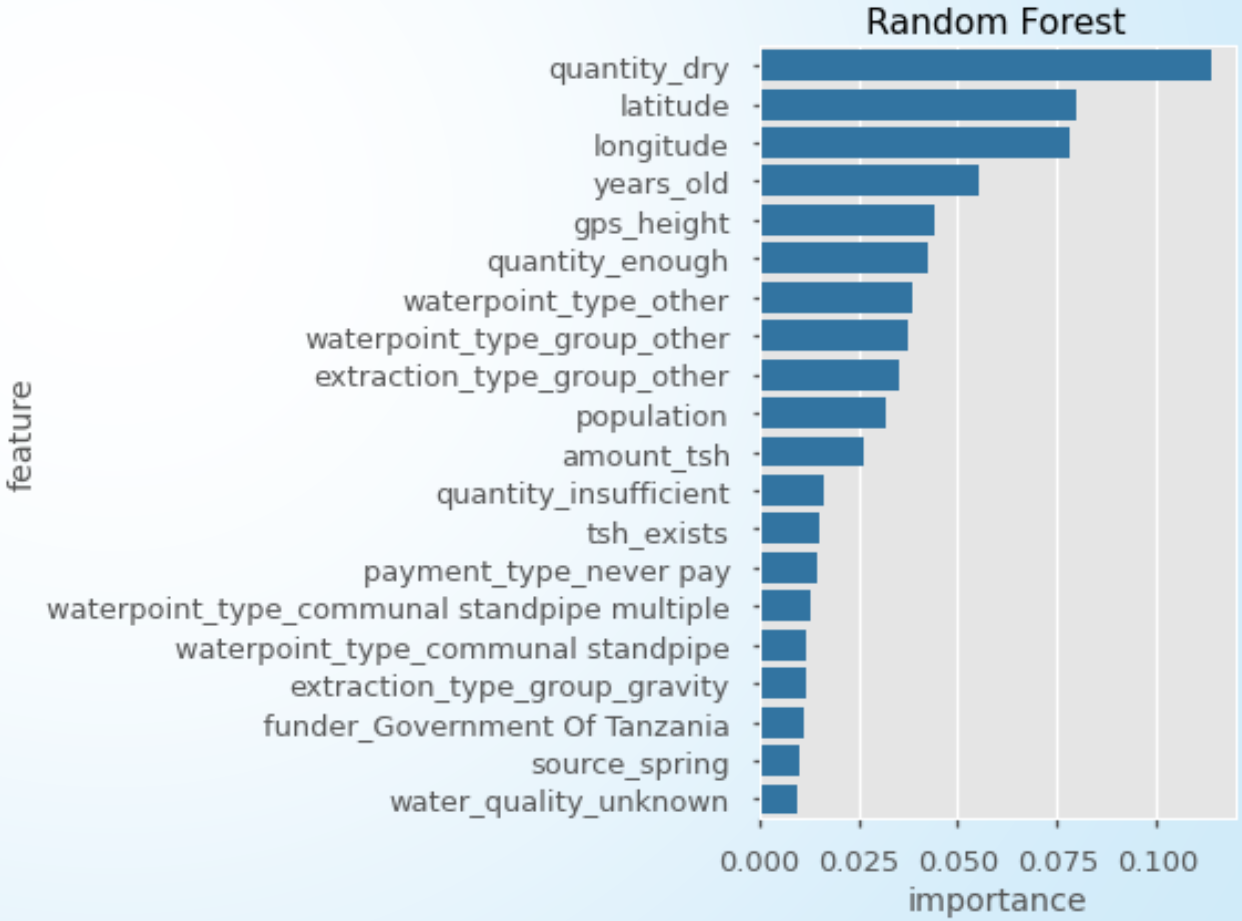
That's a huge tree!

Methodology – OSEMN



Decision Trees

Decision Trees and *Random Forests* can show the most important features that they used to make their predictions.



Methodology – OSEMN

Interpret

- One of the great advantages of a Random Forest Classifier is its **interpretability.**
- Because of its method of classification (using splits of the data), the most vital features can be easily obtained.
  - A Random Forest using ***all 137 features*** of the data scores **81% accuracy**.
  - A Random Forest using ***only the 11 most-important features*** scores **79% accuracy**.

Methodology – OSEMN

Interpret

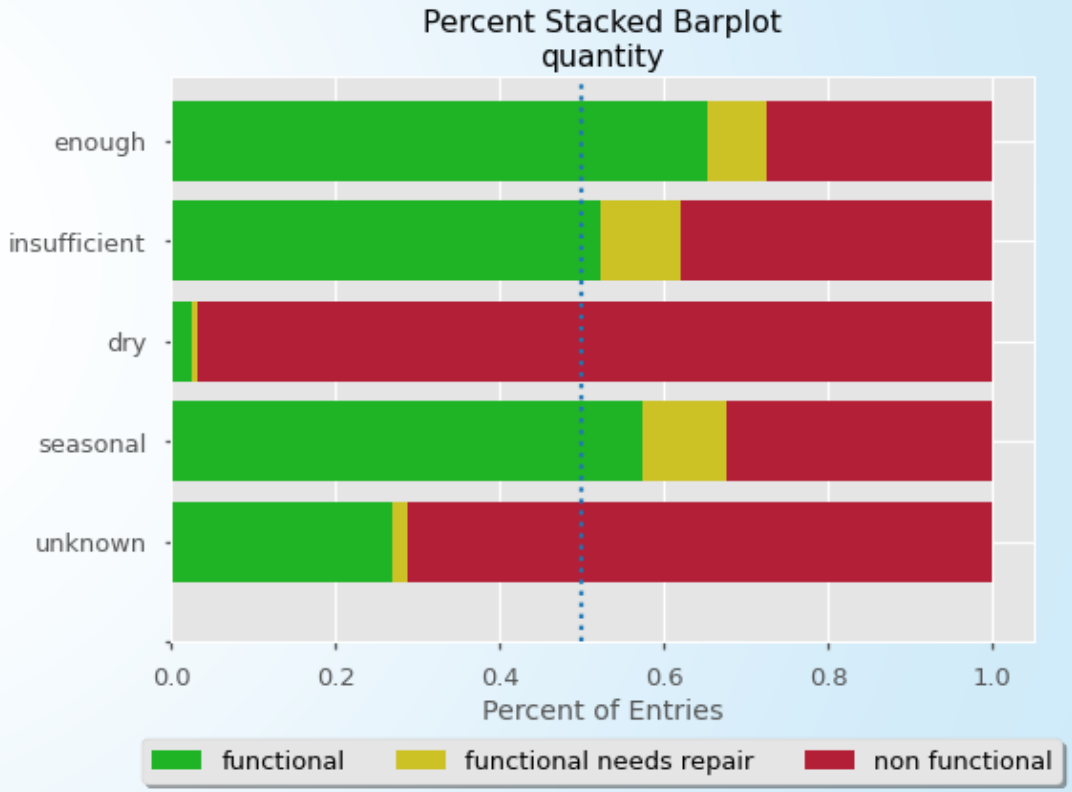
The 5 most important features for the Forest are:

1. Quantity - dry: Is the well dry or not?
2. Latitude: Geographical latitude of the well.
3. Longitude: Geographical longitude of the well.
4. Years old: How many years ago was the well installed?
5. GPS height: Altitude of the well.

How can I tell if a well is functional?

Recommendations

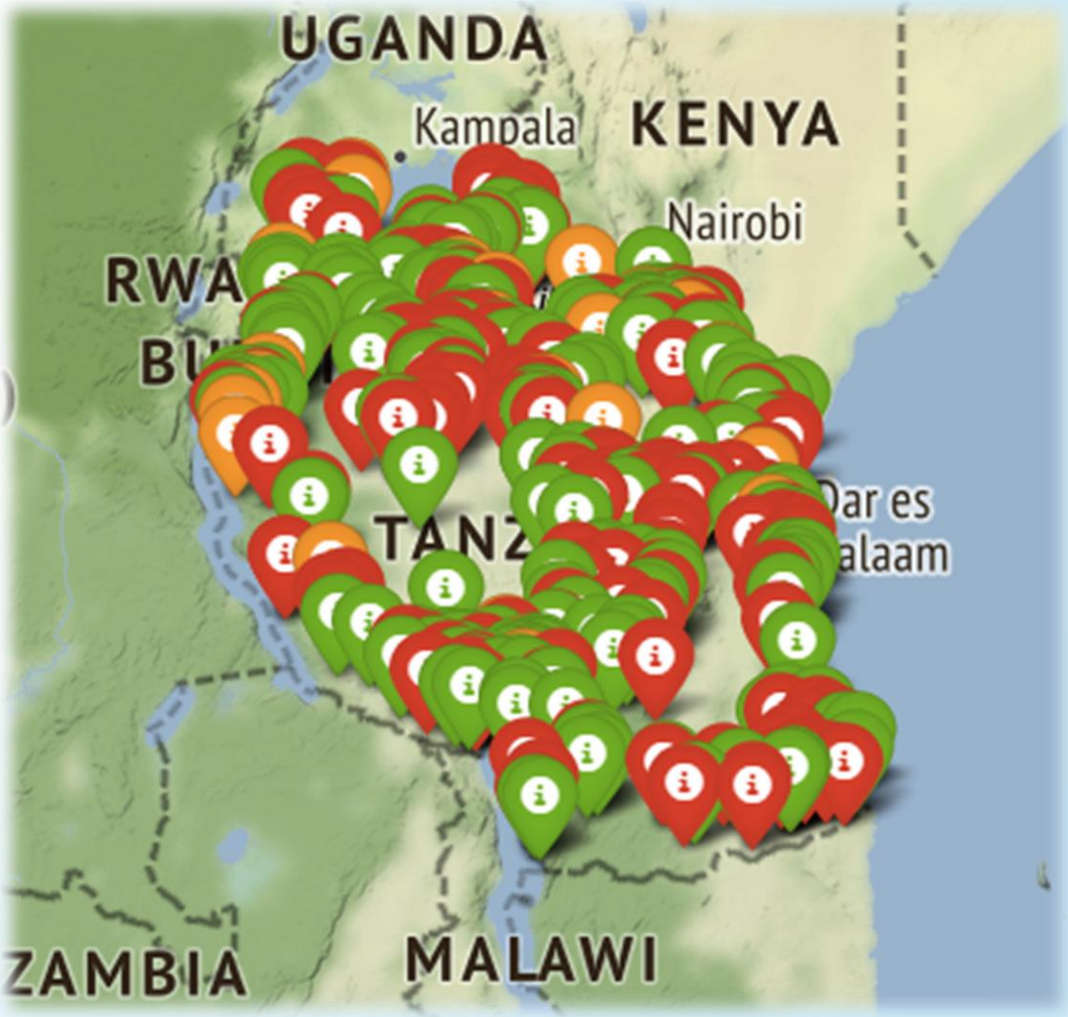
1. **Quantity**: Look for wells that with dry or unlabeled water quantities.
   - Wells that are labeled as **dry** are almost all *nonfunctional*.
   - Conversely, less than 25% of wells that have **enough** water are *nonfunctional*.
   - Note: The missing values for this feature are not missing at random (i.e.: when the label is **unknown**, it is much more likely to be *nonfunctional*).



How can I tell if a well is functional?

Recommendations

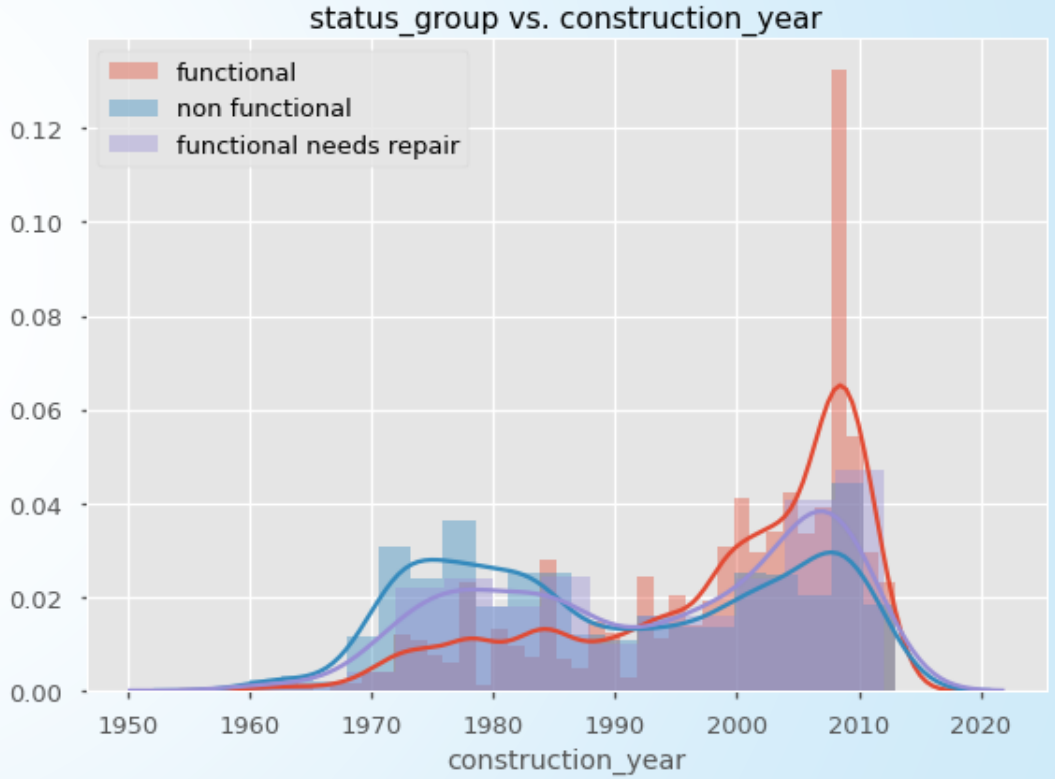
2. **Latitude / Longitude**: While there are functional and *nonfunctional* wells spread over the country, there do seem to be "pockets" of non-functional wells. This could be due to the water source or other geographical features of an area.
    - If there is a pocket of *nonfunctional* wells in an area, other wells in the same area might be *nonfunctional* as well.



How can I tell if a well is functional?

Recommendations

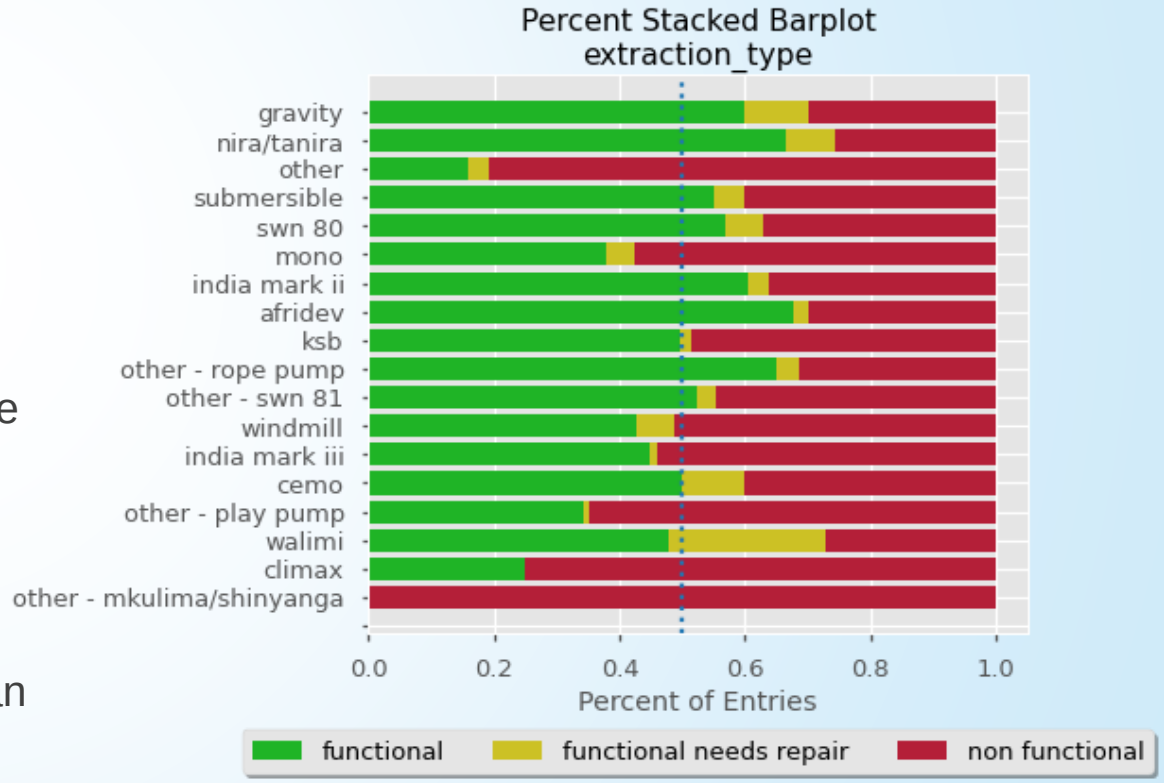
3. **Years Old**: Look for older wells (especially built before 1990).
   - As the construction year increases, the wells become much more likely to be functional.
   - Wells built before **1990** are more likely to be *nonfunctional* or *needing repair*, while after 1990 are more likely to be *functional*.



How can I tell if a well is functional?

Recommendations

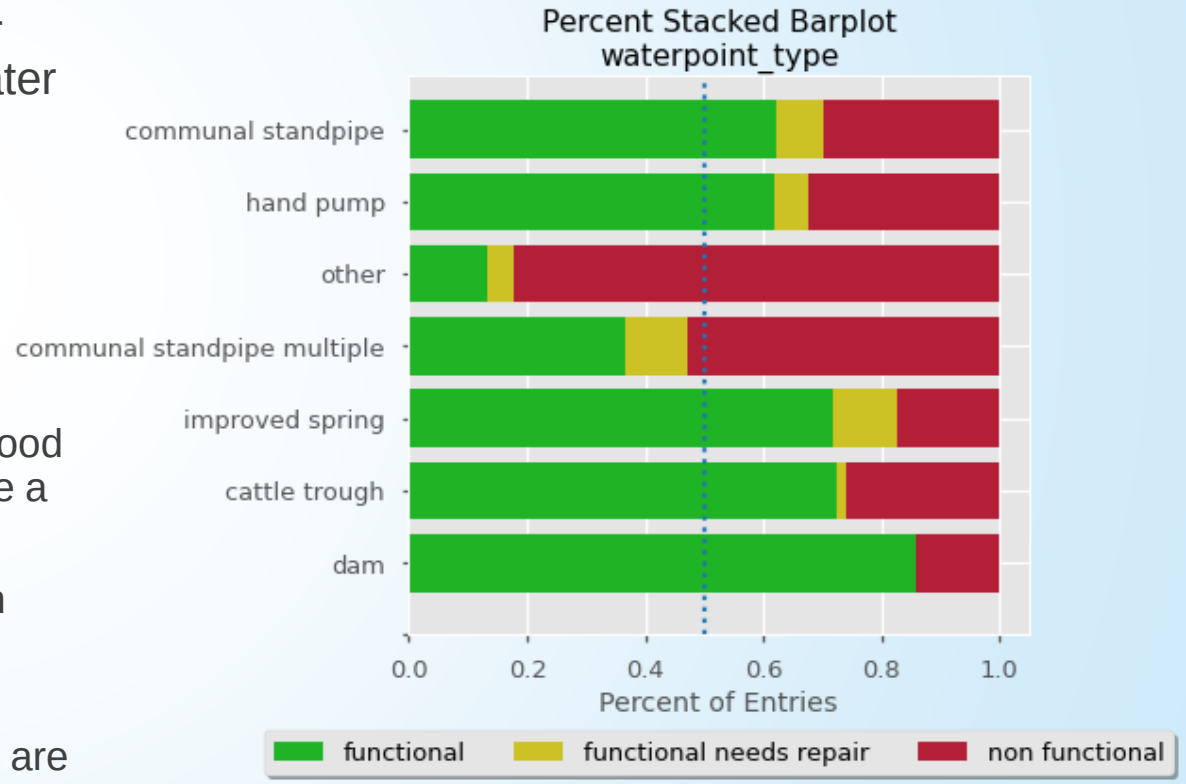
4. **Extraction Type**: Determine the extraction type.
   - **Gravity** pumps (the most common) are 60% likely to be *functional* and 25% to be *nonfunctional*.
   - The most common type or well to have more *nonfunctional* wells than *functional* is **mono**. These have over a 60% chance to be at least *in need of repair*.
   - Note: Wells that are labeled **other** have a much higher likelihood of being *nonfunctional* than otherwise.



How can I tell if a well is functional?

Recommendations

5. **Waterpoint Type**: Find wells that have less common waterpoint types (other than a communal standpipe) for greater likelihood that it will need repairs.
   - The most common **communal standpipe** is over 60% likely to be *functional* and 25% to be *nonfunctional*.
   - Having a **communal standpipe multiple** though increases the likelihood of the pump being *nonfunctional* quite a lot.
   - If the waterpoint type is less common (marked **other**), it is almost certainly (over 80%) *nonfunctional*.
   - It seems that **improved spring** wells are the most likely to need repairs (these waterpoint types may require more frequent maintenance than others).



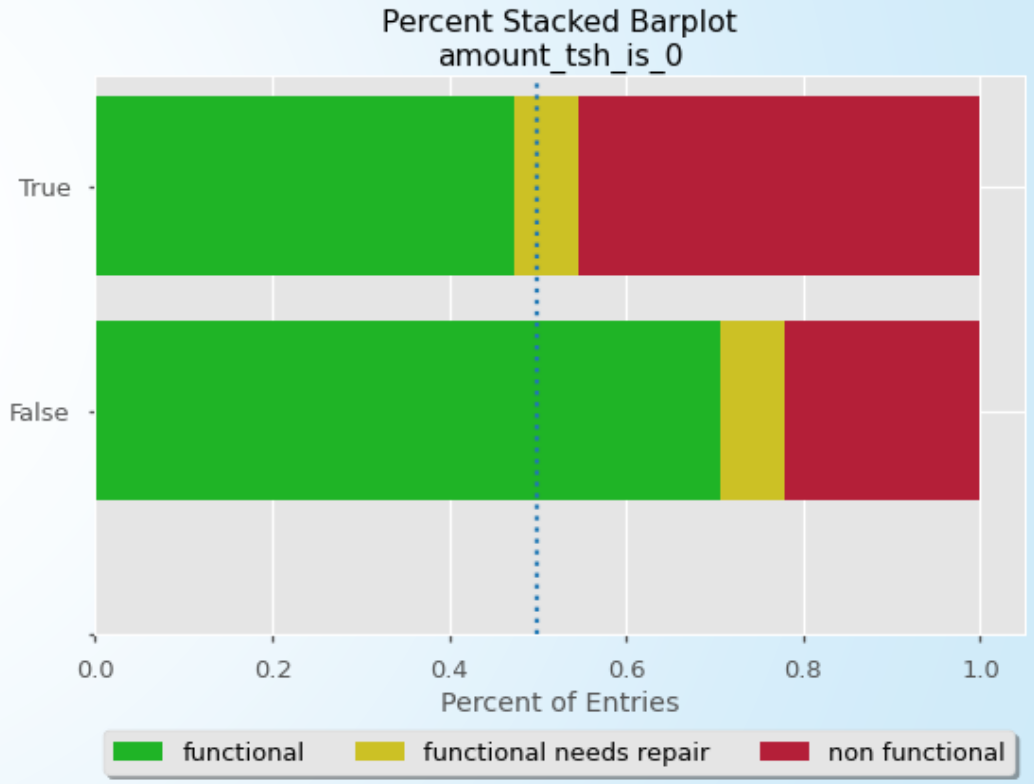
How can I tell if a well is functional?

Recommendations

6. **Amount Total Static Head**: Most wells have 0 total static head (available static water), but some have as much as 200,000 units.

   Wells that have 0 **tsh** are much more likely to be *nonfunctional* compared to wells with more than 0 **tsh**.

   - As a general principal, the greater the **tsh**, the more likely it is to be *functional*.



Next Steps

- Dive deeper into the **location** of each well.
  - Are there regions where certain well-types work best?
  - Are certain water-sources likely to run dry soon?
  - Are there geographical features (mountains, deserts, plains) that impact the wells' condition?
  - Are wells within a close vicinity to one another more likely to be nonfunctional?

Thank You!

- Flatiron School
- Taarifa (*http://taarifa.org/*) and the Tanzanian Ministry of Water.