Predicting Water Wells In Need Of Repair For The Government Of Tanzania

By: Kaelyn Isaac Parris

Overview

- Tanzania
  - Country in East Africa within the African Great Lakes region.
  - Faces challenges in providing clean water to it's population of over 57,000,000
- Many wells have been added
  - Some need repairs.
  - Some have stopped working.
- Goal:
  - Build classification model to identify broken wells for The Ministry of Water.

Business And Data Understanding

- **Data source:** Taarifa waterpoints dashboard
  - Aggregates data from the Tanzania Ministry of Water
  - Open source platform for crowd-sourced reporting and triaging of infrastructure-related issues
  - Helps engage citizens with their local government

Modeling

I WILL ATTEMPT TO ANSWER THE FOLLOWING QUESTIONS:

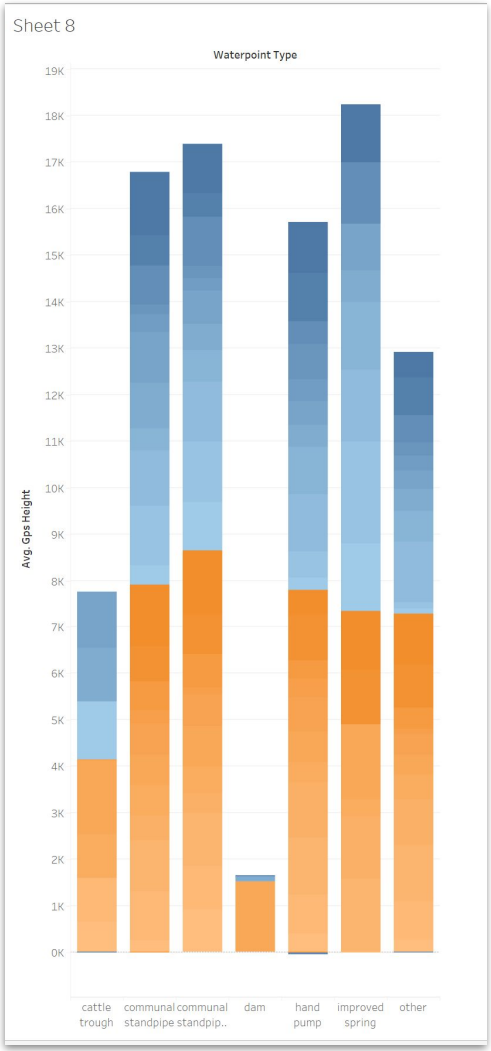
- Is there a pattern in regards to:
  - WHO?
    - Government
    - private business
    - etc
  - WHERE
    - Hotspots
    - Patterns in location
  - And PERMIT STATUS
    - Does it affect probability of repair status?

Evaluating The Model

- Prioritizing False Negatives
  - Showing up to a functioning well is worse than ignoring a broken one!
- Therefore, used Precision metric
  - Used minimize false negatives
    - Occur when the model incorrectly identifies negative instances.
  - Trade off with Accuracy
    - Will identify more false positives as a result.
- Final model precision:
  - 79%
  - Model correctly identified 79% of broken wells as positive

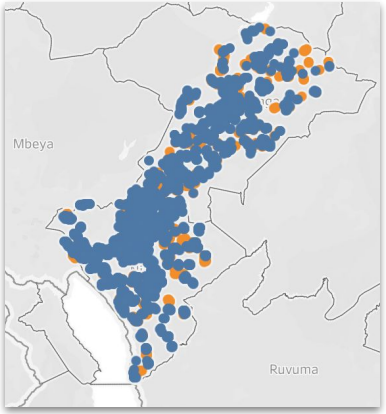
Model Results:

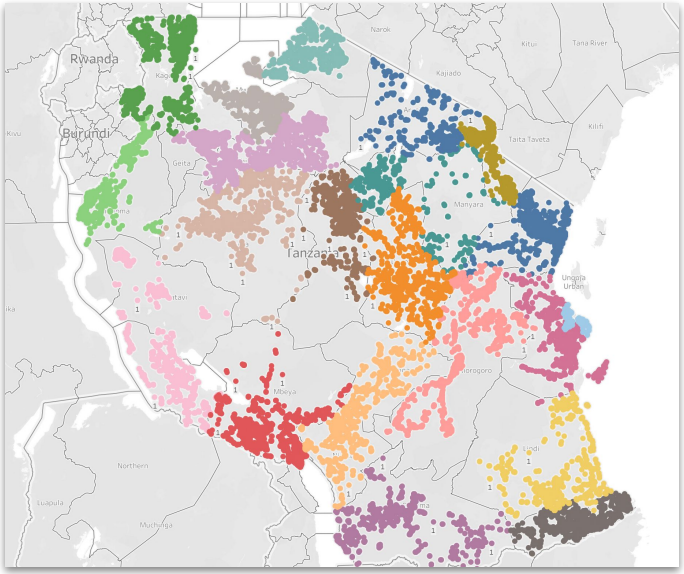
- WHO:
  - No pattern in data used.



Model Results:

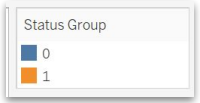
- WHERE:
  - Iringa
    - Has the highest number of broken wells.
  - Highly correlated with Longitude/Latitude
    - There are hotspots
      - But the model itself can't tell you where
  - Other Indicators:
    - High Population
    - Year Constructed

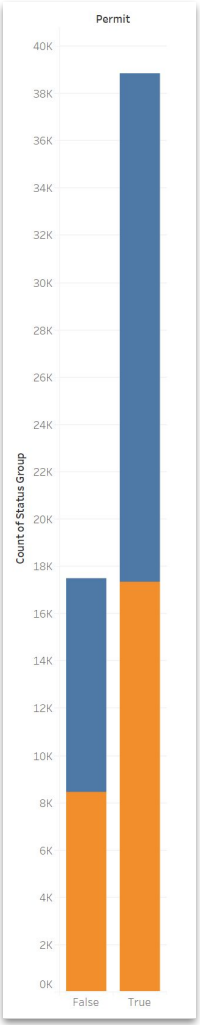




Model Results:

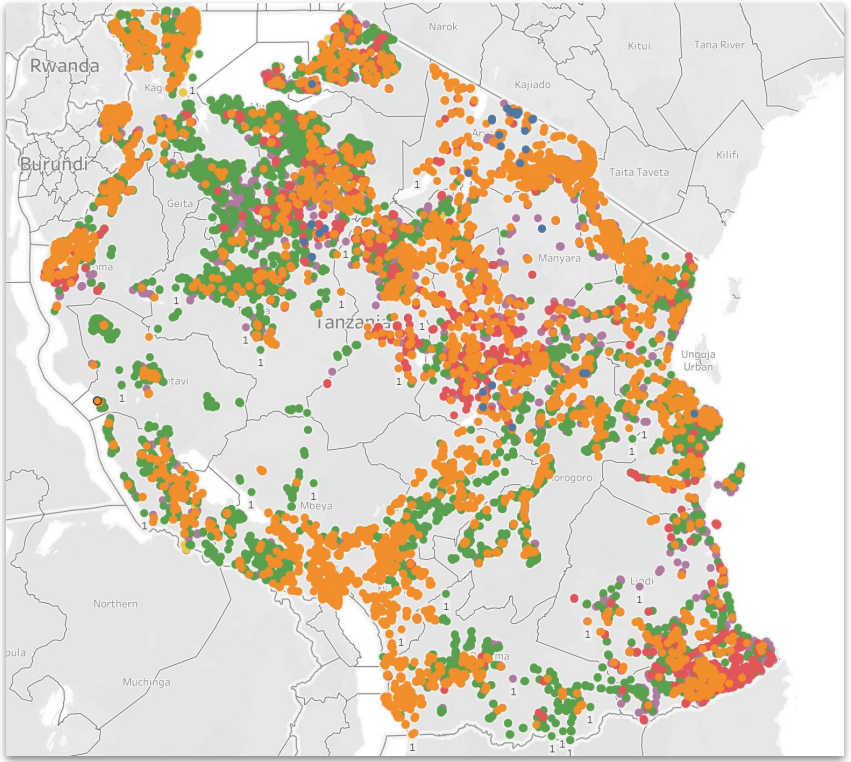
- Permits:
  - Don't seem to matter.

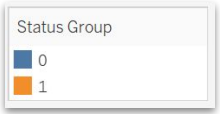


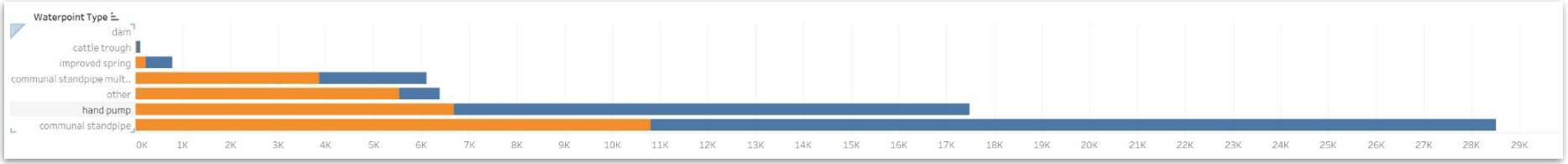


Model Results:

- Most waterpoints are
  - Communal
  - Hand Pump







Recommendations & Next Steps

- Biggest Problem
  - Data has overlapping categories
    - Could use some cleaning up.
  - Very difficult to manage multicollinearity
- Look into hotspots:
  - Some places are more likely than others.
  - Iringa
- Look into specifically
  - Communal Standing Pipes
  - Handpumps
  - Pumps with
    - Low water
    - Old
    - High Population

Questions?

Contact info:

kaelyn_parris@protonmail.com

Kaelyn Parris | LinkedIn