Tanzanian Pump It Up Water Project

Build a Python Classifier to predict the condition of a water well.

By: TJ Whipple
Flatiron School Project
March 2020

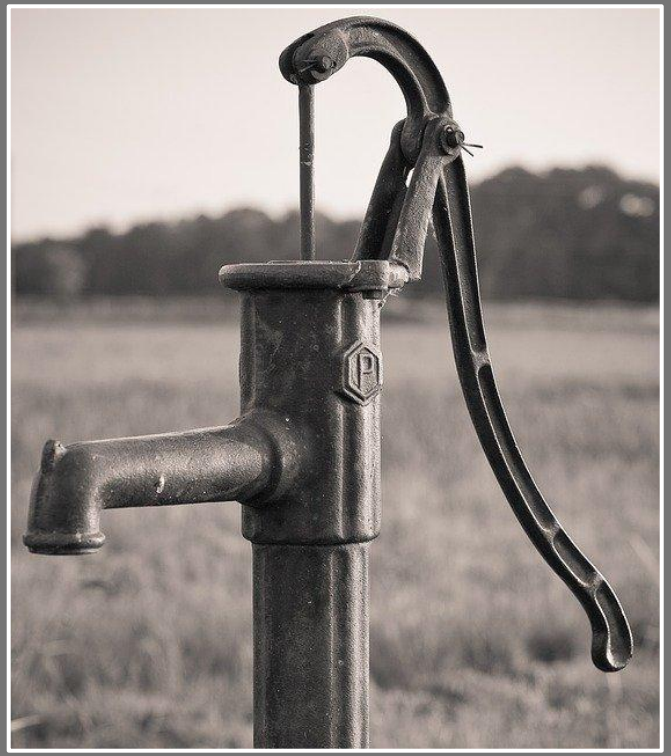
Test Methodology

A non-technical description of how the classification tests were performed.

The data was shared through the DrivenData competition website. This dataset was first cleaned and scrubbed - meaning that any missing values or duplicates were removed. The data was then randomly split into two groups - a training group and a testing group. Various machine learning algorithms were applied using Python and SciKit learn libraries to the training group. The model was then applied to the testing group. These results were analysed and compared in order to determine the best overall model.

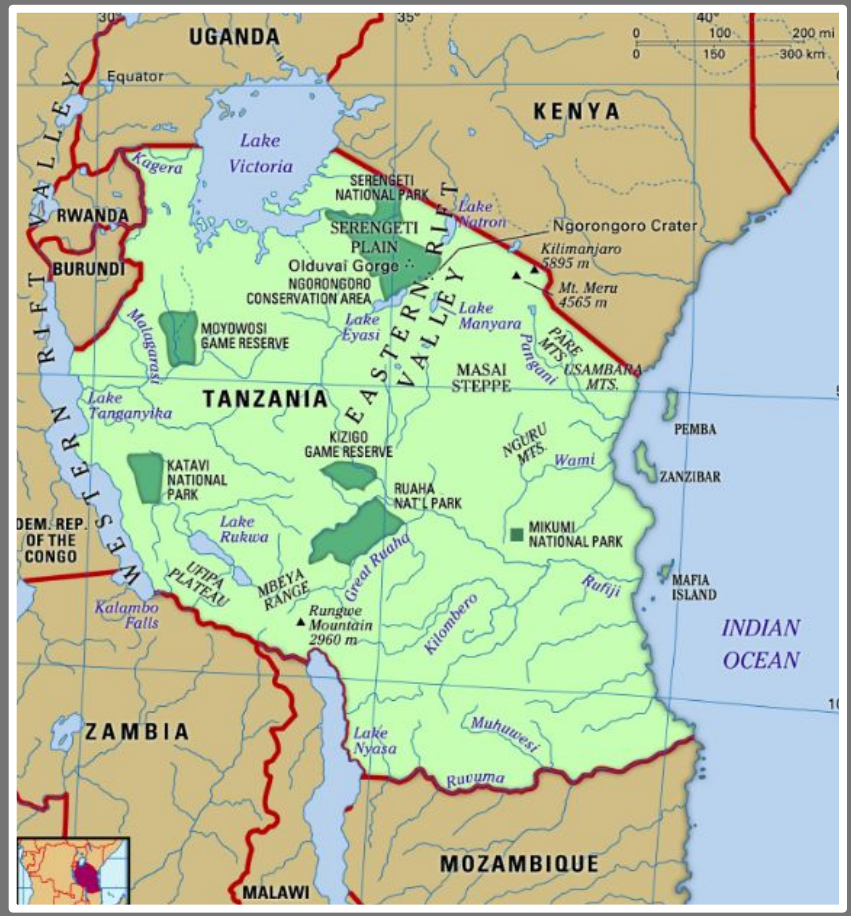
Project Focus

The focus of the project was to understand which wells, pumps, and water-points might fail or prove to be non-functional. This could help to improve maintenance operations and ensure that clean, potable water is available to the communities of Tanzania. Data was collected by the Tanzanian Ministry of Water and the Taarifa Platform.

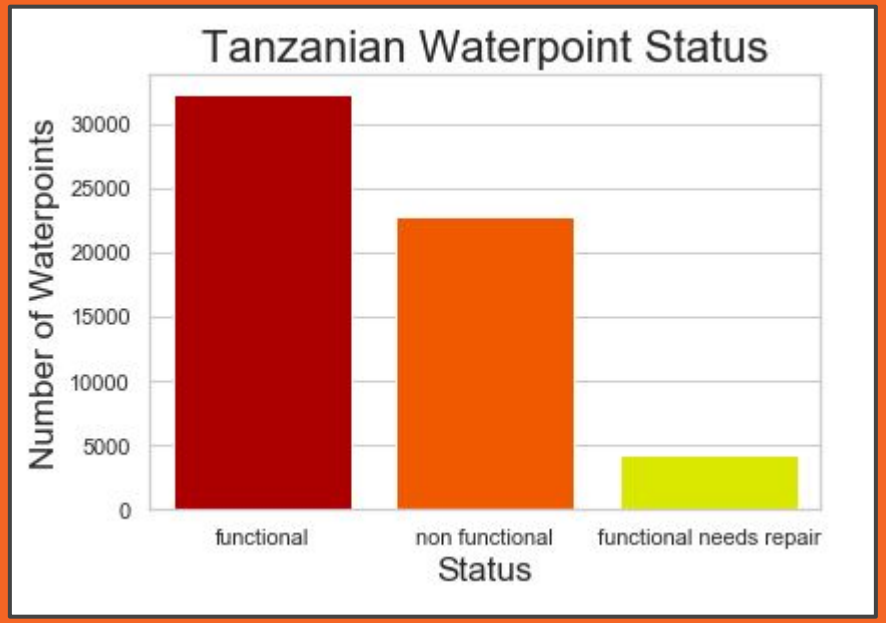


Tanzania

Population: 60 million
Capital: Dodoma
Size: About 2 times California
Location: Eastern Africa
Famous Sites:
The Serengeti National Park
Lake Victoria
Mt. Kilimanjaro - 19,341 feet
Zanzibar Island
Language: Swahili/English
Water-Points in Data: 59,900



The Three Classes

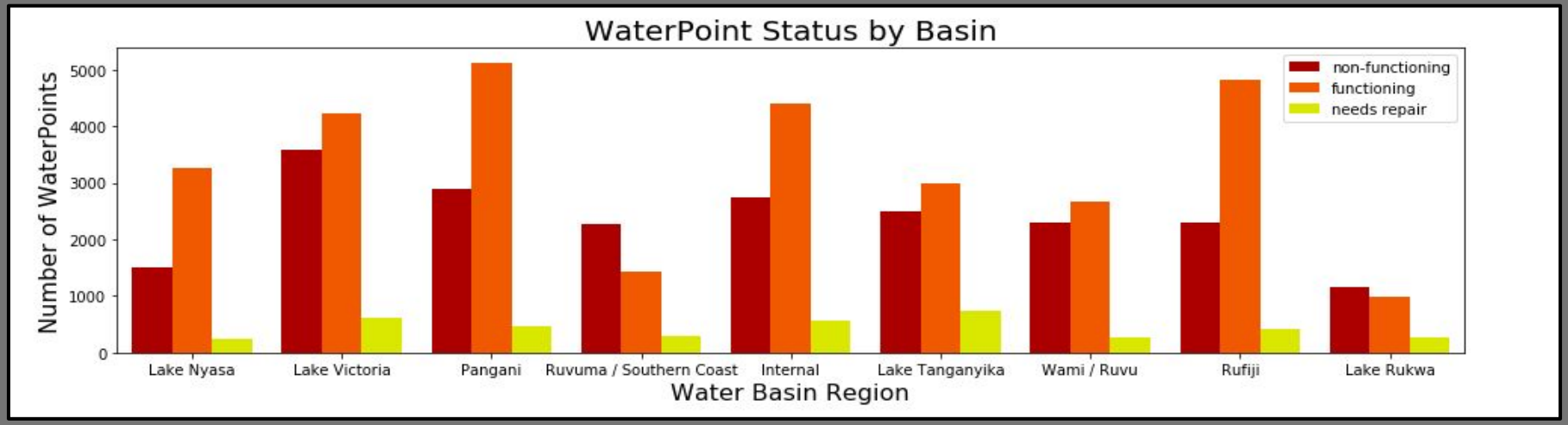


The goal was to use our data science models to predict the operating condition of a water-point for each record in the dataset.

The dataset included 59,900 water-points which were labelled as “functioning”, “non-functioning”, or “functioning needs repair”.

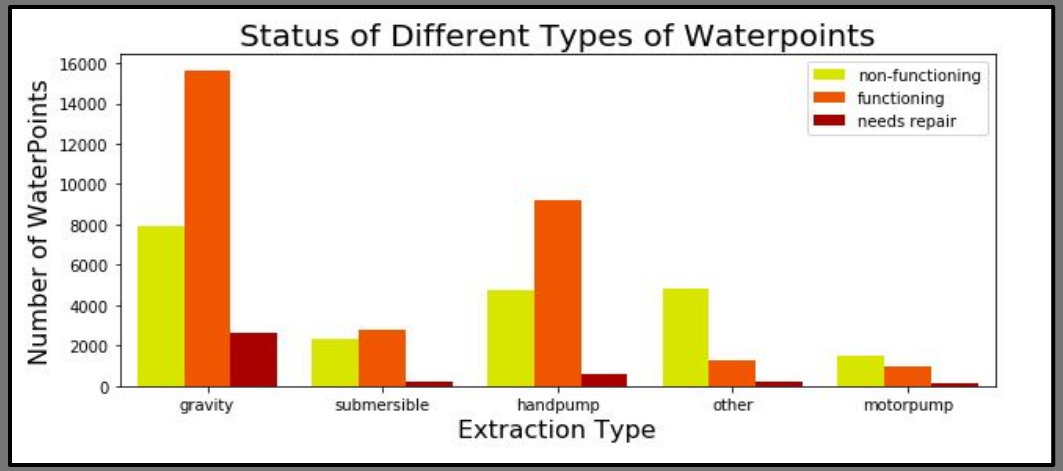
Water Basin Findings

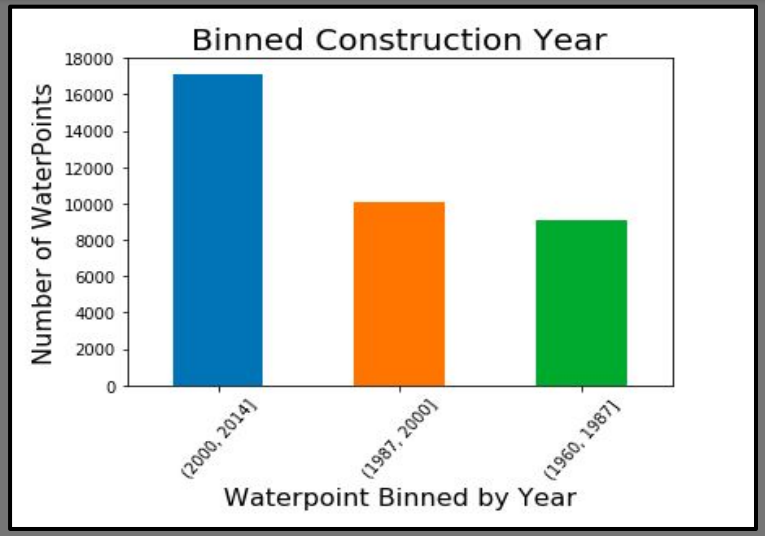
- Lake Victoria Region - most non-functional. (red bar graph)
- The Ruvuma/Southern Coast - high non-functional percent.
- Rufiji and Pangani - most functioning. (orange bar graph)
- Lake Tanganyika - most in need of repair. (yellow bar graph)



Other Results

- About 30% of the data had no entry for construction year.
- A quarter of the water-points were more than 25 years old.
- Obviously, newer pumps were better.

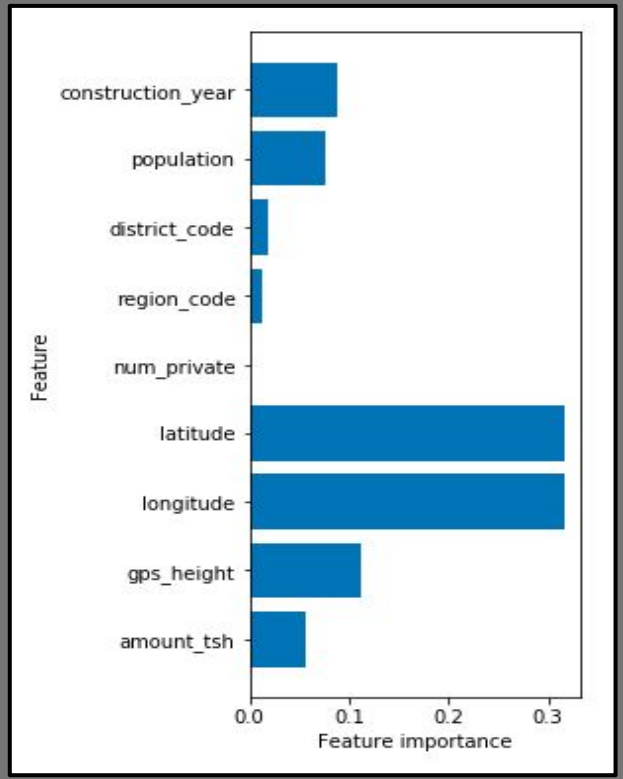




- Gravity and Hand-pumps were the most functioning.
- Other extraction types performed worse.

Summary

- The best model was able to predict water-points with 80% accuracy.
- Monthly, annually, and per bucket payments were the most efficient. Obviously, waterpoints with a payment system were less likely to need repair.
- Extraction methods of hand pumps and standpipes were the most likely to remain functional.
- Location was consider important



Future Work

- Focus on the water quality - the most important aspect of a healthy community is clean water!
- More information on number of people using pumps and number of visits is necessary!
- Any data on when a water-point was last serviced/replaced and a list of issues (i.e. broken handle, no water, power outages) would help.
- A more regulated collection of information and more precise entry of the relevant data is needed (water-point age, type, amount, etc.)

Thank you for your time.

Questions or Comments?

Sources:

- The DrivenData Pump It Up Competition
- Flatiron School
- Python, Pandas libraries, Scikit.learn, Jupyter Notebook, GitHub, Matplotlib

Presentation By: TJ Whipple

Copyright:

https://www.countryflags.com/en/tanzania-flag-image.html

https://pixabay.com/photos/water-hand-pump-water-pump-pump-1227680/

https://free-images.com/display/wishing_well_well_old.html

https://mapcruzin.com/free-tanzania-maps.htm