Predicting Water Pump Access

by Brayton Hall

The goal of this project: predict the functionality of water pumps throughout villages in Tanzania.

retrieved from water.org

The dataset: obtained from Taarifa (Rwandan News) and the Tanzanian Ministry of Water via the open prediction competition on drivendata.org

Why are water pumps important?

What are they?

Water pumps come in a variety of forms.

They are often the only source of clean water for a village.

Why do they matter?

According to WHO, nearly 800 million people lack access to safe water.

Being able to predict pump failure could help parts of the world which rely on pumps for clean water.

retrieved from: datadriven.org

An initial look at the data

Features

- 59, 400 samples
- Location, funding, pump type, nearby water basins
- **21 final features** after dropping duplicate columns or those missing excessive values

Target

- **water pump functionality**
- 'functional' or 'non-functional'
- combined 'functional needs repairs' with functional

Initial Correlations

Pump type: 0.22

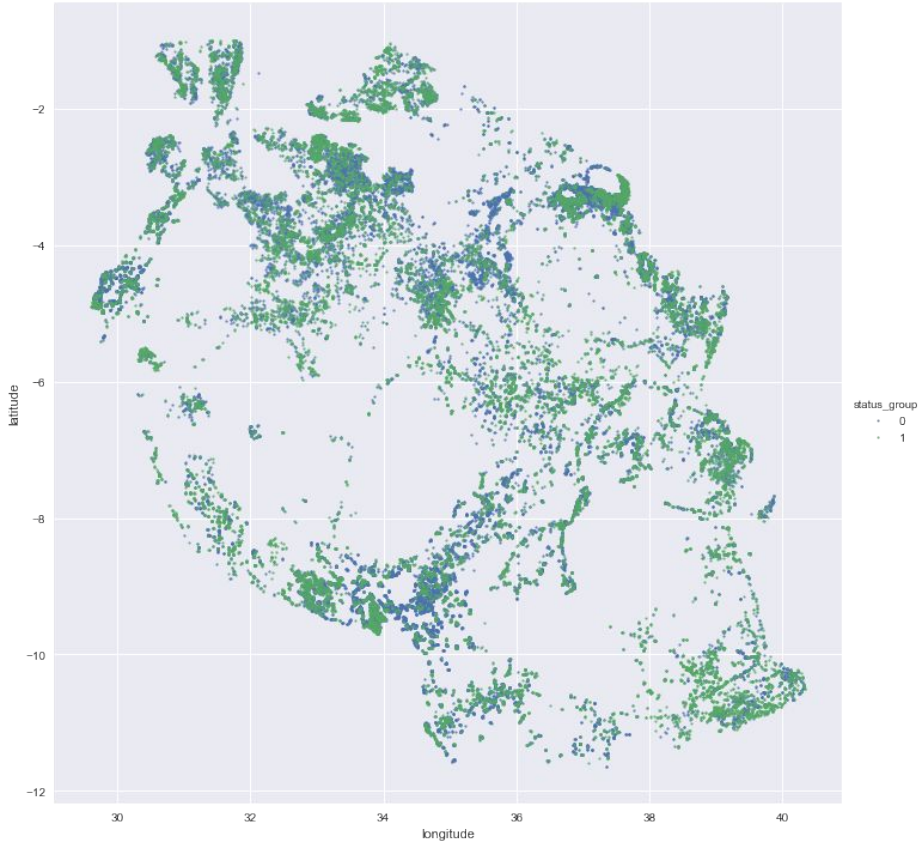
Quality: .16

Elevation: -.11

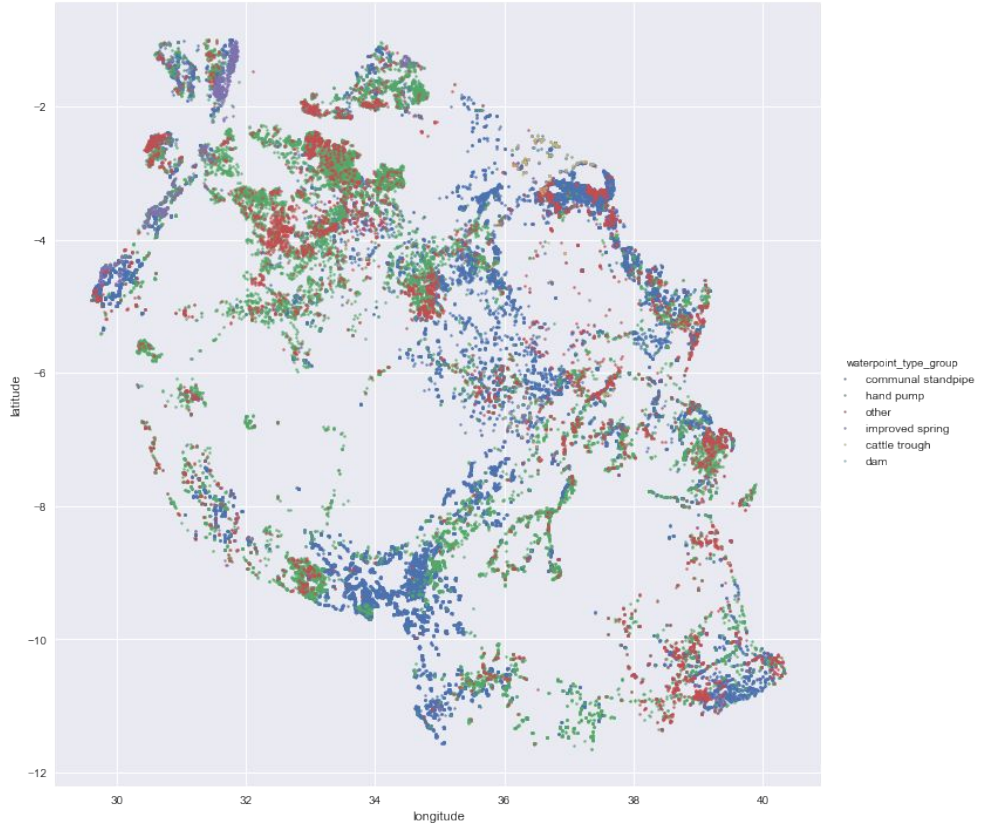
Region: .11

Pump functionality by LOCATION

Blue dots are NON-functional pumps

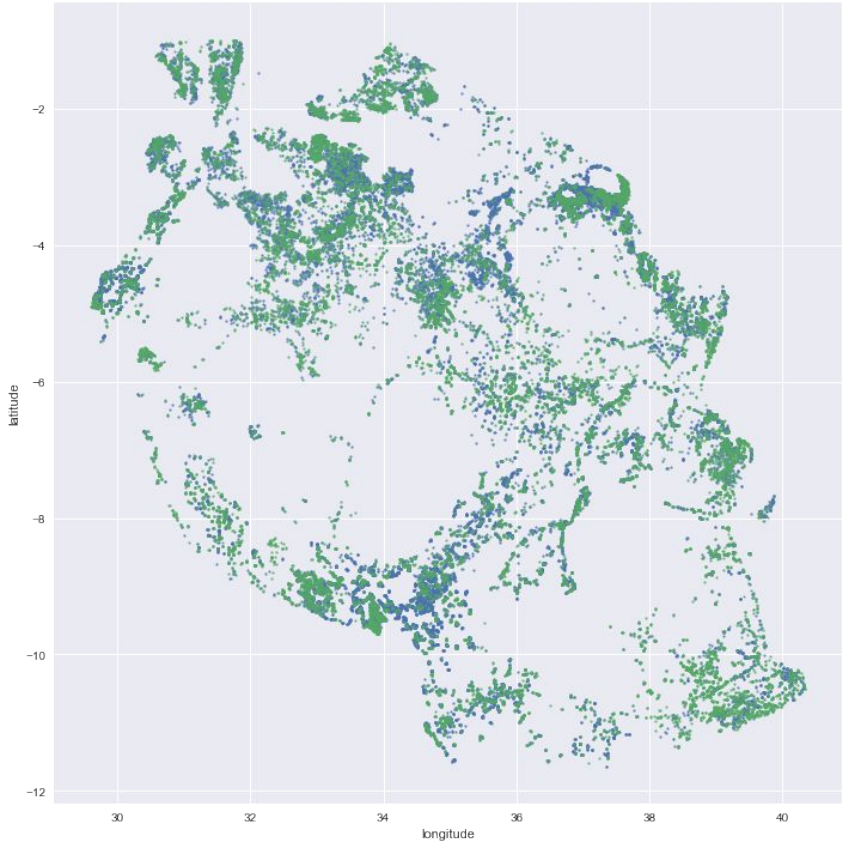


Blue dots are communal standpipes

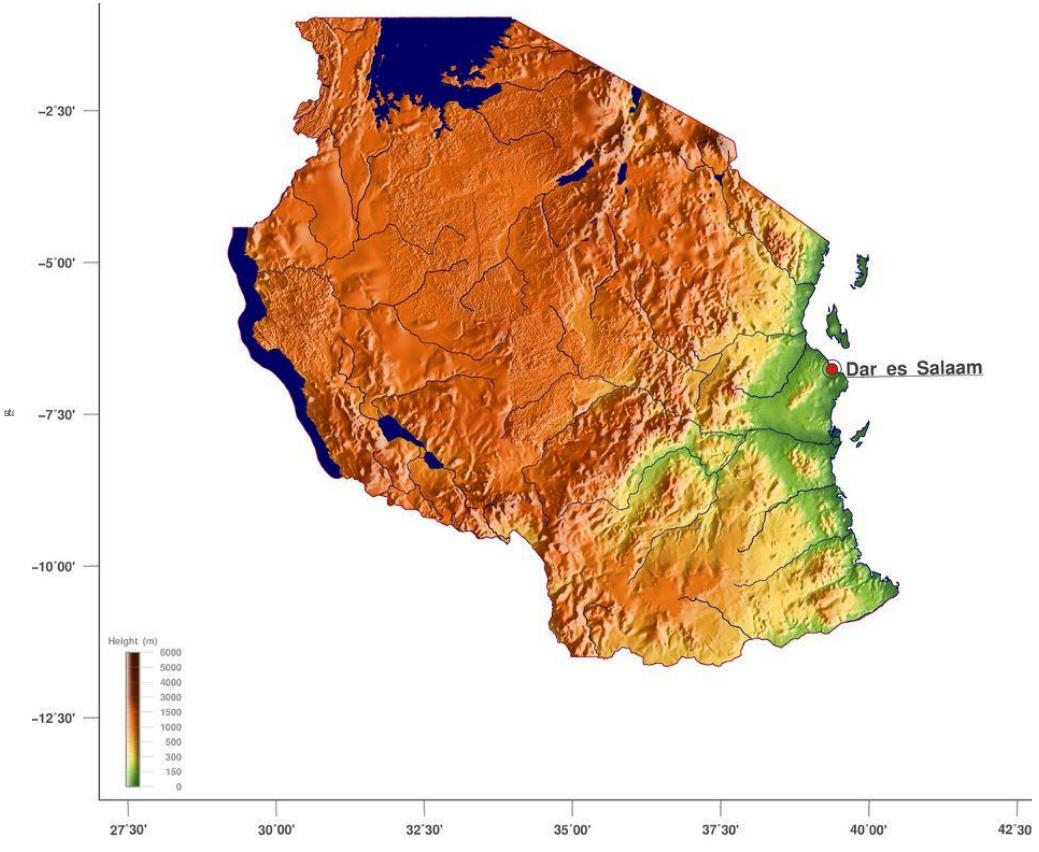


Pump functionality by ELEVATION

Blue dots are NON-functional pumps



retrieved from: mapsland.com/africa/tanzania



MODELS

Type	F1 SCORE	ACCURACY SCORE
BASELINE:	0.56	0.38
RANDOM FOREST:	0.58	0.76
KNN (7 Neighbors)	0.72	0.80
LOGISTIC	0.55	0.63
KNN GRIDSEARCH (9 Neighbors)*	0.74*	0.81*

*K-Nearest Neighbors was, surprisingly, the best model. Perhaps due to the clustering of the target variable (pump functionality) around predictive collinear features.

Looking at the most important features

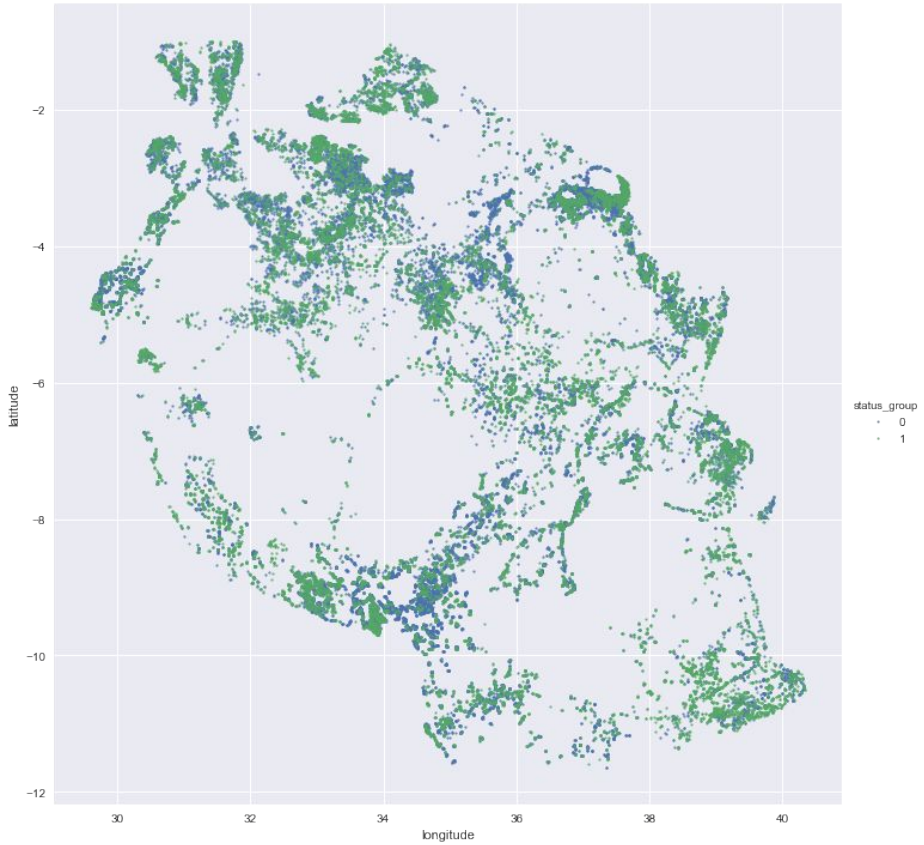
- Elevation
- Population
- Construction year
- Region

Measured using 'F-score' with xgboosts 'plot importance' function. This measures how many times a feature was split on.

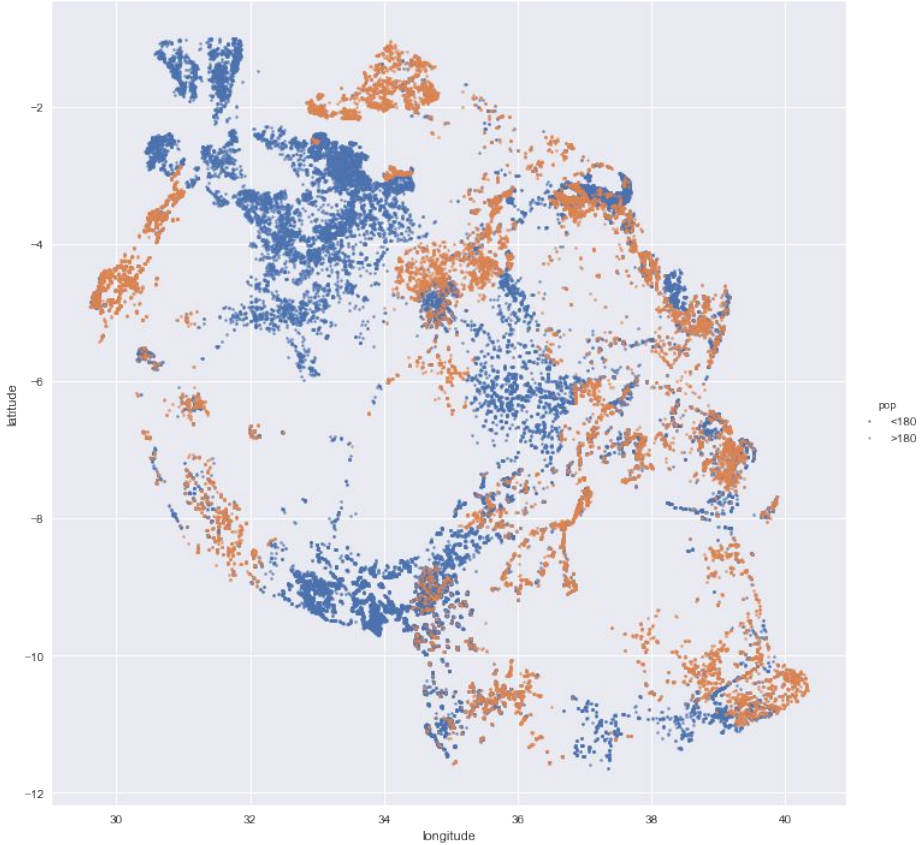
Interestingly, pump type was not selected as an important feature to split on, despite being the strongest initial correlator with the target.

Pump functionality by POPULATION

Blue dots are NON-functional pumps

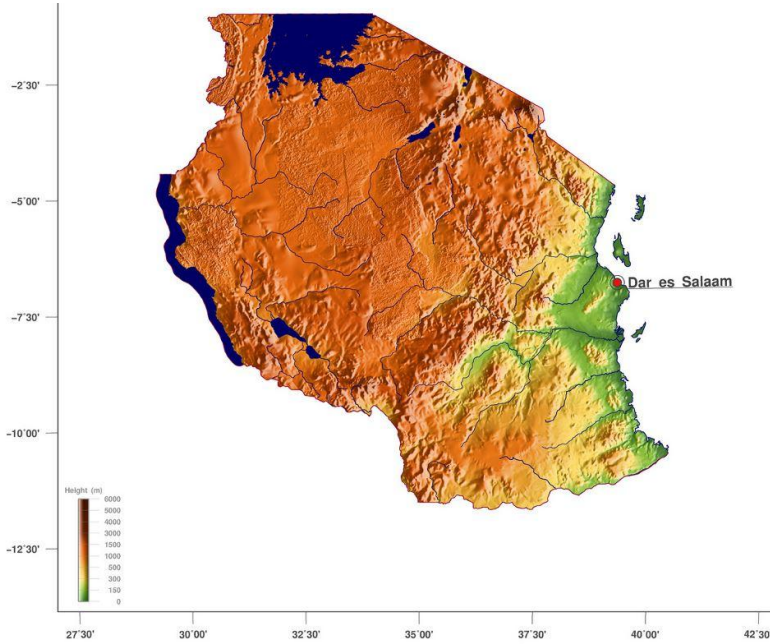


Orange dots are populations less than 180



Conclusion

Elevation, Population, Construction year, and Region were the most important features for prediction pump functionality.



These aren't easily actionable findings, since they indicate a **broader systematic problem** involving overuse by isolated mountain villages without resources to repair or build new pumps.

Solutions would most likely require providing isolated communities with material assistance.