Pump it Up: Data Mining the Water Table

BY FILIP SZAFRANSKI

Can you predict which water pumps are faulty?

An attempt to predict which pumps are functional, which need some repairs, and which don't work at all using data from Taarifa and the Tanzanian Ministry of Water.

What is the problem?

Over 24 million people are impacted by the The United Republic of Tanzania's water crisis; that's almost half of the population of Tanzania

A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

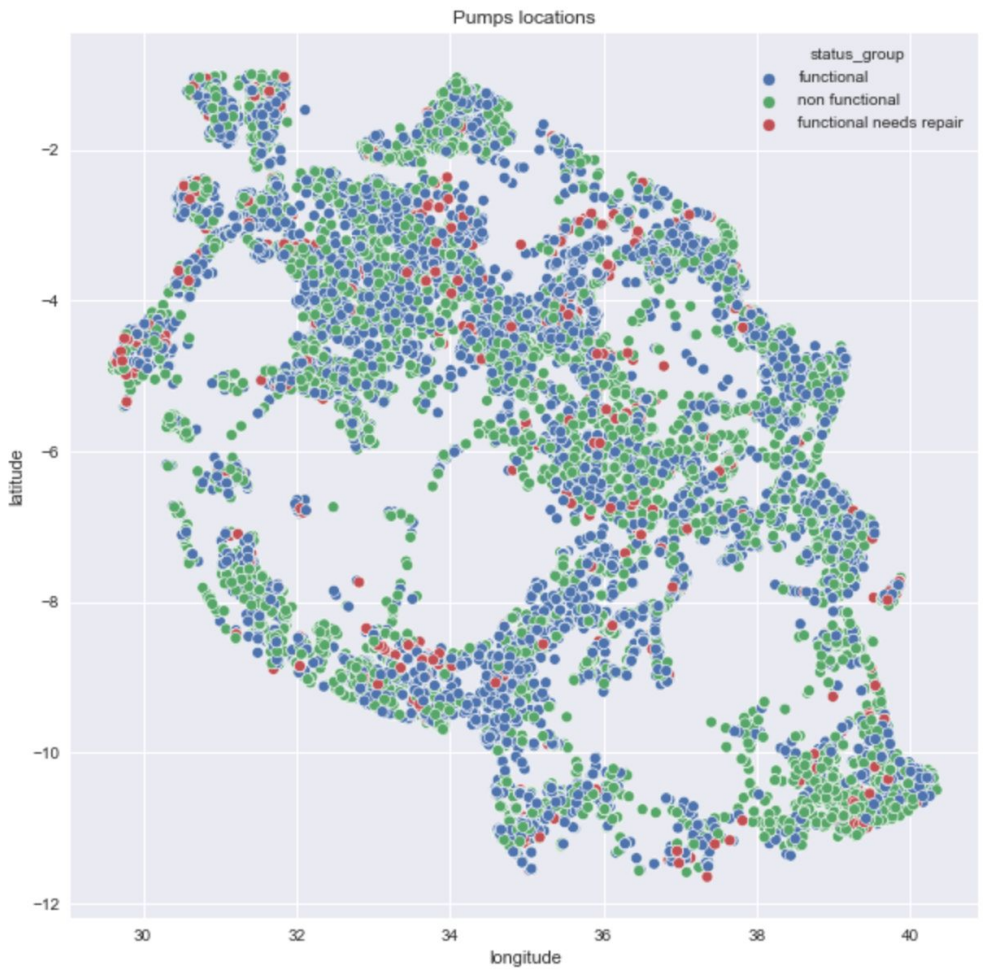
The goal is to predict one of these three classes based on a number of variables about what kind of pump is operating, when it was installed, and how it is managed.

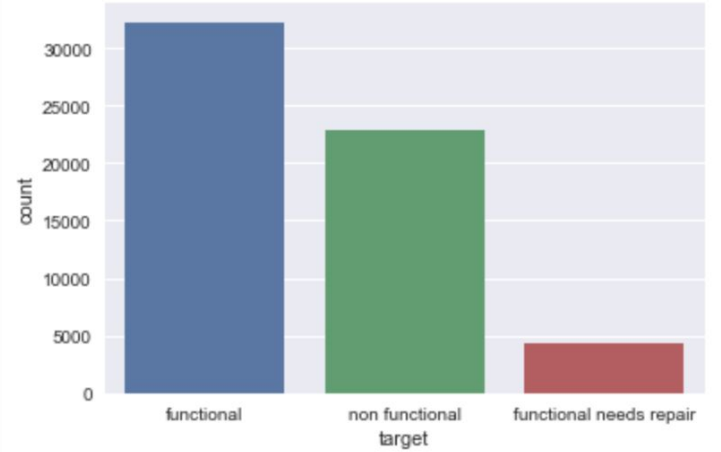
The Data.

Data from Taarifa and the Tanzanian Ministry of Water provides almost **60.000** records of water pumps across Tanzania.

The majority of the pumps are functions however almost **30% of the pums is not functional.**

From left: Water pumps locations indications functionality status, number of water pumps in each functionality class





The Data.

There is the following **set of information** about the waterpoints provided in the dataset. Only the some, relevant variables were listed.

- amount_tsh - Total static head (amount water available to waterpoint)
- date_recorded - The date the row was entered
- funder - Who funded the well
- gps_height - Altitude of the well
- installer - Organization that installed the well
- longitude - GPS coordinate
- latitude - GPS coordinate
- wpt_name - Name of the waterpoint if there is one
- region - Geographic location
- region_code - Geographic location (coded)
- lga - Geographic location
- population - Population around the well
- public_meeting - True/False
- scheme_name - Who operates the waterpoint
- construction_year - Year the waterpoint was constructed

	NaN count	Zero values count	Unique_val count	Data type
scheme_name	28166	0	2697	object
scheme_management	3877	0	13	object
installer	3655	0	2146	object
funder	3635	0	1898	object
public_meeting	3334	5055	3	object
permit	3056	17492	3	object
subvillage	371	0	19288	object
num_private	0	58643	65	int64
amount_tsh	0	41639	98	float64
population	0	21381	1049	int64
construction_year	0	20709	55	int64
gps_height	0	20438	2428	int64
district_code	0	23	20	int64

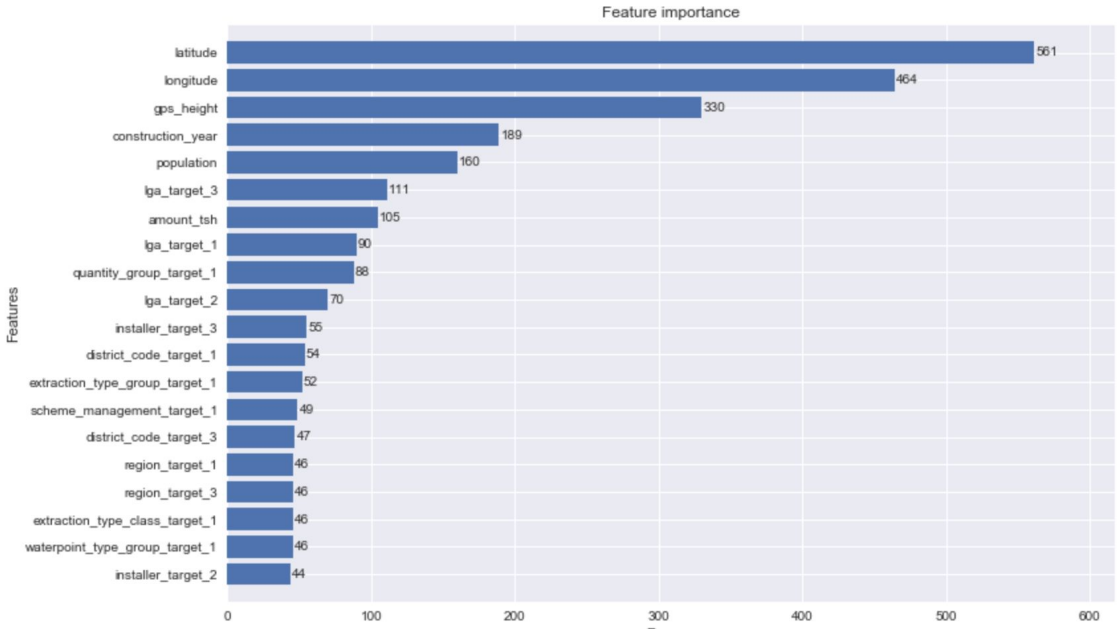
Above is the **summary** with the missing information and zero values in each category. Some of the data were replaced with the median of the values in the respective category or if the number of missing values was too large, the category was removed from the dataset

The Classification Model.

After the model was train with the provided data and validated the accuracy of the model was oscillating around **70%.**

The algorithm used to make predictions was the **XGBoost**.

XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.




```
[[4572  996  110]
 [ 533 7384  181]
 [ 111  341  622]]
```

Above confusion matrix showing correctly predicted values

	precision	recall	f1-score	support
0	0.65	0.79	0.71	5678
1	0.79	0.72	0.75	8098
2	0.45	0.18	0.25	1074
accuracy			0.71	14850
macro avg	0.63	0.56	0.57	14850
weighted avg	0.71	0.71	0.70	14850