Pump it Up: Data Mining the Water Table

First conclusions from the data analysis

# Presentation of the challenge:

Using data from Taarifa and the Tanzanian Ministry of Water, can you predict which pumps are functional, which need some repairs, and which do not work at all? This is an intermediate-level practice competition. Predict one of these **three classes** based on several variables about what kind of pump is operating, when it was installed, and how it is managed. 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 features in this dataset

Your goal is to predict the operating condition of a waterpoint for each record in the dataset. You are provided the following set of information about the waterpoints:

* 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
* num\_private -
* basin - Geographic water basin
* subvillage - Geographic location
* region - Geographic location
* region\_code - Geographic location (coded)
* district\_code - Geographic location (coded)
* lga - Geographic location
* ward - Geographic location
* population - Population around the well
* public\_meeting - True/False
* recorded\_by - Group entering this row of data
* scheme\_management - Who operates the waterpoint
* scheme\_name - Who operates the waterpoint
* permit - If the waterpoint is permitted
* construction\_year - Year the waterpoint was constructed
* extraction\_type - The kind of extraction the waterpoint uses
* extraction\_type\_group - The kind of extraction the waterpoint uses
* extraction\_type\_class - The kind of extraction the waterpoint uses
* management - How the waterpoint is managed
* management\_group - How the waterpoint is managed
* payment - What the water costs
* payment\_type - What the water costs
* water\_quality - The quality of the water
* quality\_group - The quality of the water
* quantity - The quantity of water
* quantity\_group - The quantity of water
* source - The source of the water
* source\_type - The source of the water
* source\_class - The source of the water
* waterpoint\_type - The kind of waterpoint
* waterpoint\_type\_group - The kind of waterpoint

# Distribution of Labels

The labels in this dataset are simple. There are three possible values:

* functional - the waterpoint is operational and there are no repairs needed
* functional needs repair - the waterpoint is operational, but needs repairs
* non functional - the waterpoint is not operational

# Submision Format:

id,status\_group

50785,functional

51630,functional

17168,functional

45559,functional

...

# Training set analysis:

The training set if made of 59400 observations. Only 2% of the data is missing, and 36 rows are duplicated. In total, 41 columns, including 29 categorical, 9 numerical and 2 Boolean.

A big pre-processing of the data is needed, to encode categorical variables.

* Duplicate rows will be deleted

### Columns to drop:

* Amount\_tsh: 70% missing values -> mapped to 1 if >2e4
* Recorded by: constant value set to 59400
* Funder: cardinality to high (1900) -> reduce to 9
* Installer: cardinality to high (1900) -> reduce to 9
* wpt\_name
* lga
* Num\_private:
* Subvillage: cardinality to high (19287)
* Region\_code: correlated to region
* District\_code
* Ward
* Scheme\_name
* Extraction\_type: correlated to extraction\_type\_class
* Extraction\_type\_group: correlated to extraction\_type\_class
* Management: correlated to management\_group
* Payment\_type: correlated to payment
* Water\_quality: we keep quality\_group (very correlated)
* Quantity\_group: same as quality
* Source: we keep source\_type
* Source\_class: correlated to source\_type
* Waterpoint\_type: we keep waterpoint\_type\_group (correlated)
* Population: 26% missing values
* Public\_meeting
* Date\_recorded
* Management\_group

### Missing values:

* Funder: ”Unknown”
* Installer: “DWE”
* Scheme\_management: VWC
* Construction\_year: 1994
* Longitude: Mean ?
* Permit: False

Categorical variables to one-hot encode:

* Basin
* Scheme\_management
* Extraction\_type\_class
* Payment
* Quality\_group
* Quantity
* Source\_type
* Waterpoint\_type\_group
* installer

Numerical values:

* Gps\_height: [0, 1]
* Longitude : [0, 1]
* Latitude : [0, 1]
* Construction\_year : [0, 1]