Pump it Up: Data Mining the Water Table

- **User**: Jiang Li

URL: https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/

Goal

We will use Python/Tensorflow to do ML analysis

- **Exploratory Data Analysis(EDA)**: Looking at features which will include
  - What is the missing value for each of the feature
  - What is the correlations between each feature
- **Transform the data**: Given EDA result, transform the feature data to make it work for the ML Pipeline
- **ML Analysis**: use the feature to predict whether a **Pump** is functional or not

Challenge Summary

Can you predict which water pumps are faulty?

Using data from Taarifa and the Tanzanian Ministry of Water, can you predict which pumps are functional, which need some repairs, and which don't work at all? This is an intermediate-level practice competition. 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. 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

Figure 1:

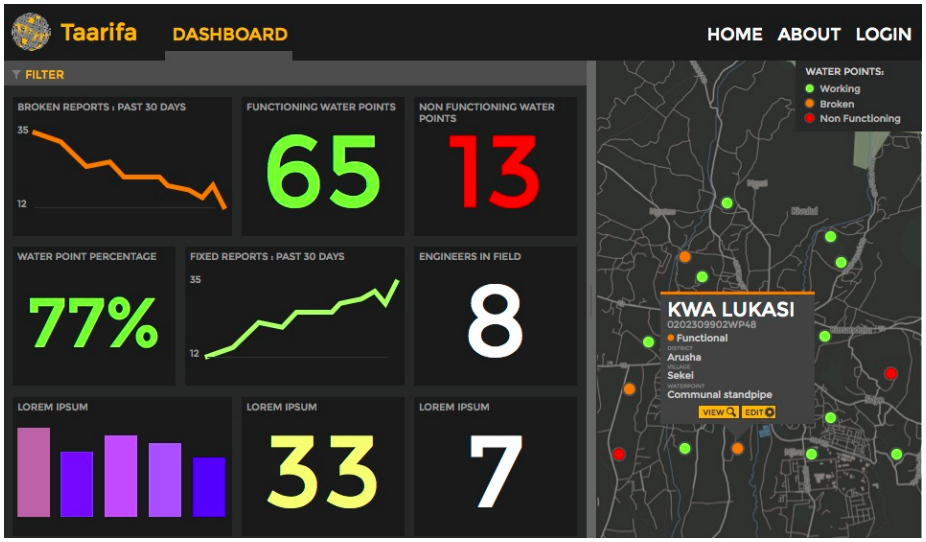

Figure 2:

- **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

The labels in this dataset

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

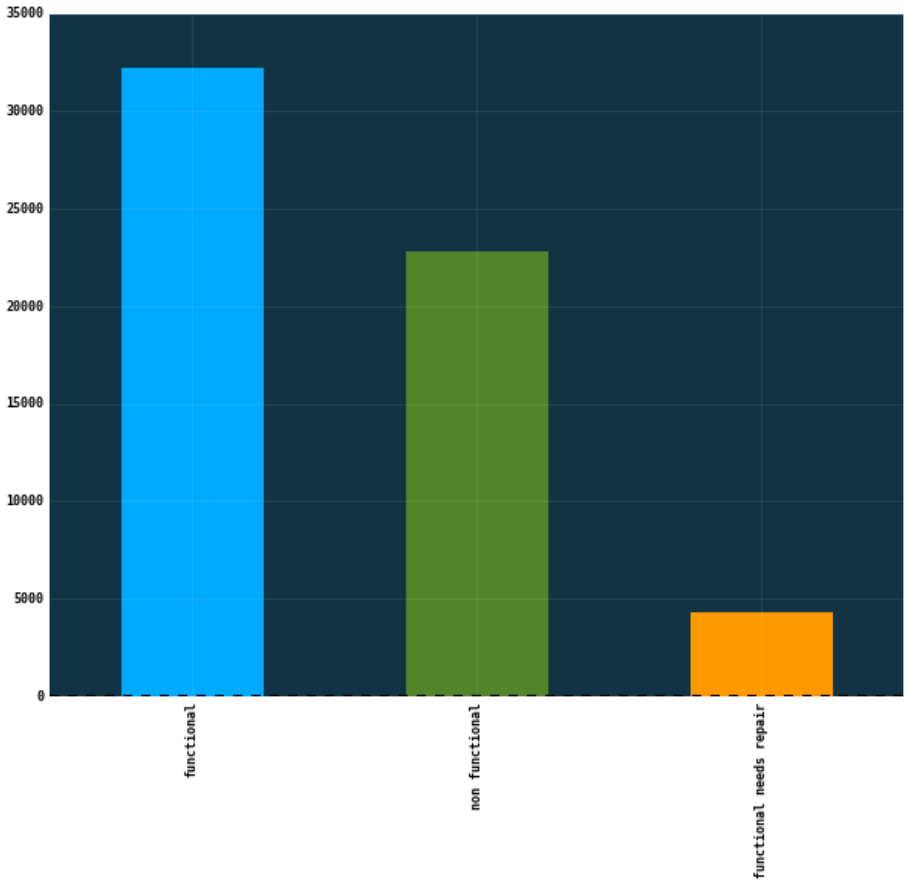
35000
30000
25000
20000
15000
10000
5000
0
functional
non functional
functional needs repair

Figure 3:

Submission format

The format for the submission file is simply the row id and the predicted label (for an example, see **SubmissionFormat.csv** on the data download page.

Your .csv file that you submit would look like:

```
id,status_group
50785,functional
51630,functional
17168,functional
45559,functional
...
```