Pump It Up: Mining the Water Table

Project Goals

Daniel Beutler
Austin Harrison
Frances Carr
Tine Hutchinson

April 22, 2018

Submitted to:
Dr. Donald Wedding
Chief Executive Officer
DAFT PUMPS
donald.wedding@northwestern.edu

Contents

DAFT PUMPS
REPAIR, PREVENTION & ANALYTICS

April 22, 2018

Dr. Donald Wedding
Chief Executive Officer
DAFT PUMPS

RE: Pump it Up: Data Mining the Water Table - Project Goals

Dear Dr. Wedding,

Thank you for the consideration of DAFT PUMPS to assist the Tanzanian Ministry of Water with the classification and prediction of the operational condition of water pumps based on the Ministry's data as well as Taarifa. Please find attached DAFT PUMPS' project goals documentation. We hope that this gives you a complete description of our insight into the project, preliminary plan, overall goals and intended deliverables throughout this process.

Please do not hesitate to reach out to us if you need further clarification at any time.

Sincerely,

Daniel Beutler
Austin Harrison
Frances Carr
Tine Hutchinson

1 Executive Summary

DAFT PUMPS intends to work with the data provided by Driven Data to assess and classify the data to accurately predict the operational condition of various waterpoints (or pumps) in the throughout Tanzanian region using the following labels: functional, functional needs repair and non-functional. An educated understanding on previous classifications in order to predict future and additional waterpoint operations will help to provide both optimized processes as well as continuous clean water to the included communities throughout Tanzania.

Previous research has been conducted on this issue and DAFT PUMPS intends to build upon and improve the classification results. For example, Darmatasia & Arymurthy (2016) used XGBoost to help predict the status of water pumps in the region with an 80% accuracy, and discussed a number of other studies that suggested the use of Artificial Neural Networks (ANN) and Support Vector Machines (SVM). The study was conducted on what appears to be the same data set. Additionally, Jiménez & Pérez-Foguet (2011) referenced that an estimated 30-46% of the Tanzanian water pumps were non-functioning.

Optimizing classification of pumps will consist of providing a consolidated view of the most up-to-date publicly available information regarding this project to build models more effectively. The predictive variables consist of a total of 39 columns and an ID column. There is information about the physical location of the wells (latitude, longitude, altitude, region, district, ward, local population), information about the building of the well (funder, installer, whether there was a public planning meeting or not, whether it was built with a permit or not, the year of construction), and the operation of the well (the management and its structure, the price charged, the amount of water, the quality of the water, the source of the water, how the water is extracted).

The project deliverables for this project include group report outs to discuss progress, the included project goals, initial findings, project report and oral report. The goals of this team are to:

1. We will provide a consolidated view of the most up-to-date publicly available information regarding this project, the data set, and prior work performed by other teams. This will allow us to build models more effectively by maximizing the value of our work. It will also position us to best speak to not just our process and model, but the distinctive value that we will provide.

2. We will score within the top 10% of participants in DrivenData's "Pump It Up: Data mining the water table" competition to show the analytics prowess we can provide in this partnership. We will accomplish this by pairing what we learned as part of the review in goal 1 with our own insights when modeling the data.

3. We will deliver a service-focused dashboard to show how we plan to utilize the model to direct our resources if we win the bid.

Overall, DAFT PUMPS plans to offer the Tanzania Ministry of Water additional and

exceptional results to the classification of water pumps and their future. We will offer consistent and reliable communication throughout the project.

2 The Problem

The purpose of this project is to classify and predict the operational condition of various waterpoints (or pumps) throughout Tanzania as either "Functional", "Functional needs repair", or "Non-Functional" (see Table 1 for more information on these classifications). By leveraging previous classifications, we will seek to predict future waterpoint status and seek methods for optimizing waterpoint operations. This will help deliver continuous clean water to people and communities throughout Tanzania.

Table 1: Label Titles & Descriptions for Waterpoint Classifications

Label	Description
Functional	Waterpoint is operational and there are no repairs needed
Functional needs repair	Waterpoint is operational but needs repairs
Non-functional	Waterpoint is not operational

Our predictions will be based on a number of variables (discussed below). Everything from well funding, location and altitude, to the cost, quality and source of the water. Additionally, we'll verify the effectiveness of our predictions with an independently evaluated test data set.

Access to water is essential to survival. As such, it is not surprising that previous research has been conducted to assist Tanzania with water pump access and assessment. Darmatasia & Arymurthy (2016) earlier used a machine learning algorithm known as "XGBoost" to help predict the status of water pumps in the region with an 80% accuracy, and discussed a number of other studies that suggested the use of Artificial Neural Networks (ANN) and Support Vector Machines (SVM). The study was conducted on what appears to be the same data set.

According to research from Jiménez & Pérez-Foguet (2011), an estimated 30-46% of the Tanzanian water pumps were non-functioning. From what we've seen so far in the data, not much has changed in the intervening years. According to the team:

> "In the first five years of operation, about 30% of water points become non-functional. Only between 35% and 47% of water points are working 15 years after installation, depending on the technology. By categories, hand pumps are the less durable of the technologies studied. We suggest that more emphasis has to be placed on the creation of community capacities to manage the services during and after the installation of water points." (Jiménez & Pérez-Foguet, 2011, p. 948)

Jiménez & Pérez-Foguet (2011) also suggest the use of water point mapping (WPM), and include a quote from Welle (2005):

> "An exercise whereby the geographical positions of all improved water points (WP) in an area are gathered in addition to management, technical and demographical information. This information is collected using GPS and a questionnaire located at each improved water point. The data is entered into a geographical information system and then correlated with available demographic, administrative, and physical data. The information is displayed using digital maps." (Welle, 2005)

Our overarching goal for this project is to build upon previous research in order to optimize water pump categorization and improve the processes, predictions and repairs of the water pumps to better improve water access and availability.

3 The Data

The data we are provided consists of a training set, with associated expected responses, and a testing set. We will build our models on the training set, make predictions based on the testing set, and submit our predictions to a 3rd party validation service that will let us know how well we did.

The predictive variables consist of a total of 39 columns and an ID column. There is information about the physical location of the wells (latitude, longitude, altitude, region, district, ward, local population), information about the building of the well (funder, installer, whether there was a public planning meeting or not, whether it was built with a permit or not, the year of construction), and the operation of the well (the management and its structure, the price charged, the amount of water, the quality of the water, the source of the water, how the water is extracted).

The training dataset consists of 59,400 entries. There are not a tremendous amount of completely missing data points, with the exception of the 'scheme_name' column, which is only has 31,234 non-missing values. Every other row with some missing data still has more than 55,000 rows. Missing is one thing we don't have to worry about too much with the training dataset, the same can not be said for data that doesn't seem to make sense. For instance, at least one well appears to be at latitude 0 and longitude 0, which is thousands of miles to the west, in the Atlantic Ocean. It's also the case that the testing data has many more missing data. As such, we'll examine each variable to determine the best approach to filling in any missing or out of range data so that we can be sure that every variable is usable for our predictions.

Many of the non-numeric variables have only a handful of possible values. We'll convert these variables to be categorical in the system ('category' in pandas, 'factor' in R, etc.). Other non-numeric variables have hundreds or thousands of possible values. What we find with these is that there are some values that repeat more than others. For these types of

variables, we'll create categories for the top 20 possible values and then create an "other" type that will hold values that don't repeat as often. Note that top 20 is arbitrary and we may expand or contract this number as needed to get the best results.

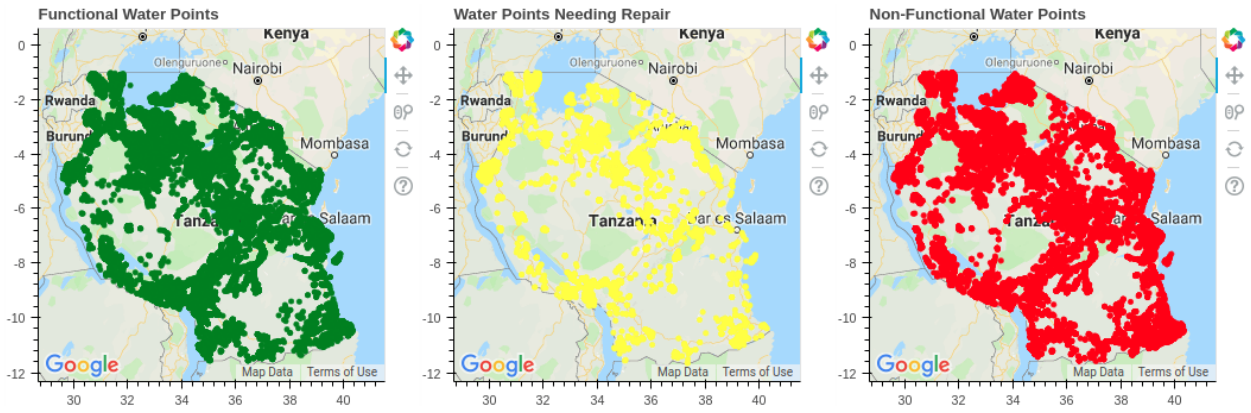
3.1 Preliminary Exploratory Data Analysis

The goal is to predict if a well is functioning, in need of repair, or not functioning. So how many wells fall into each category?

Table 2: Target Variable Distribution

Status	Functional	Functional needs repair	Non-functioning
Count	32,259	4,317	22,824
Percentage	54.31%	7.27%	38.42%

Figure 1: Geographical distribution of Water Points, by functional level



We expect that the altitude of a Water Point may be relevant to the functional status. Figure 2 gives a quick look at the "gps_height" variable, which has maybe more "0" entries than it actually should.

We have identified that construction_year, population and gps_height as typically either all being provided with values, or are all set to zero. This suggests that the construction_year was either missing or was not filled out at the time. This likely missing data is primarily distributed by region, shown in Figure 3, red represents pumps with 0 for all 3 values across these fields, orange represents 1 non-zero value, yellow represents 2 non-zero values and green represents 3 non-zero values. Only about 5% of the data has 1 or 2 values, and is not completely either 0 or non-zero. Our initial investigation indicates that models that can incorporate this data when available perform better suggesting that imputing of this data is likely important to model performance.

Figure 2: Height distribution for all wells

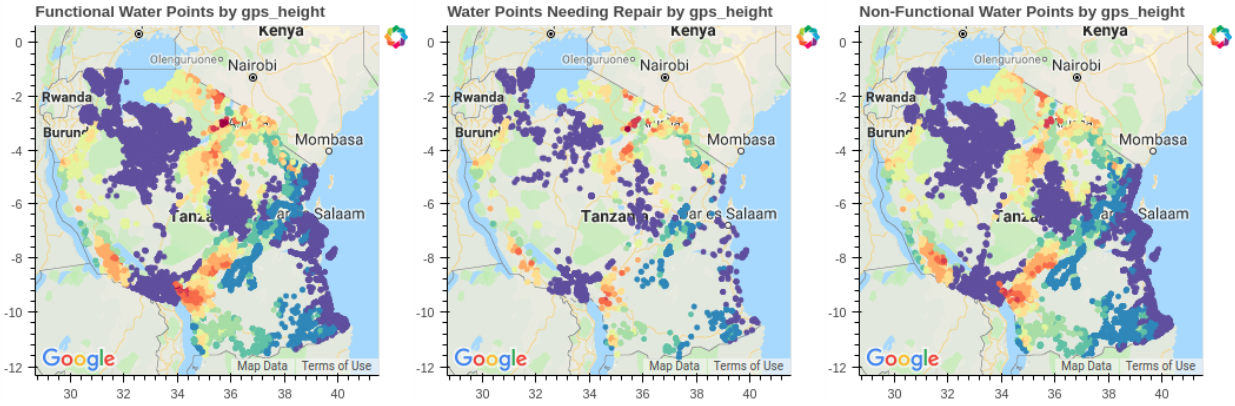
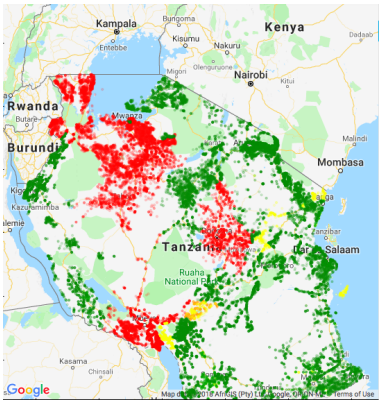


Figure 3: Missing values for construction_year, population and gps_height



4 The Tools

There are many platforms and/or languages available that can be used to perform the analysis and predictive modeling. The platforms and languages all differ in their implementations but most of them have similar capabilities with regards to modeling techniques and data exploration.

4.1 Exploratory Data Analysis

Exploratory data analysis involves examining the provided data to identify important aspects that will affect predictive models, including outliers and missing values. Often, visualizations are used to gain insights into the data and there are many options to consider when creating these visualizations. The options considered for use in this project are Python (using the packages matplotlib, seaborn, bokeh and plot.ly), R (using ggplot2), Tableau, and Power BI.

Thus far in the preliminary review of the data, Python and R have been most heavily used,

with the map-based visualizations being generated with the bokeh package for Python. Due to the flexibility provided by the Python and R programming languages, custom dashboards can be built to very detailed specifications. Another advantage Python and R have over other options is that they are free and open-source, while many of the important features in Tableau and Power BI require expensive licenses to use.

4.2 Predictive Modeling

Though selecting the right modeling tools for the job are important, the goals of the project do not include comparing the various options. Once the options have been evaluated, the team will focus on building and improving a predictive model in the chosen language or platform.

Tools considered for the predictive modeling in this project include: ANGOSS, SAS Enterprise Miner, Azure, scikit-learn and tensorflow packages in Python and caret and randomForest libraries in R. ANGOSS and Enterprise Miner have the advantage of having a graphical user interface that makes setting parameters and visualizing the flow of data manipulation and prediction. Python and R have the advantage of the availability of extensive packages and libraries that make them very flexible.

As with data exploration, Python and R will most likely be the primary tools used to train and test the predictive models. Preliminary predictive models have been created with the randomForest library in R and show promising results. Another area the team would like to explore is the usage of deep learning methods.

5 The Deliverables

5.1 Goals

In order to best improve our chances of our bid being selected this project will deliver 3 primary goals:

1. We will provide a consolidated view of the most up-to-date publicly available information regarding this project, the data set, and prior work performed by other teams. This will allow us to build models more effectively by maximizing the value of our work. It will also position us to best speak to not just our process and model, but the distinctive value that we will provide.

2. We will score within the top 10% of participants in DrivenData's Pump It Up: Data mining the water table competition to show the analytics prowess we can provide in this partnership. We will accomplish this by pairing what we learned as part of the review in goal 1 with our own insights when modeling the data.

3. We will deliver a service-focused dashboard to show how we plan to utilize the model to direct our resources if we win the bid.

5.2 Proposed Schedule

Currently we intend to deliver according to the schedule outlined below:

Week	Week Start	Week End	EDA	Modeling	Evaluation	Dashboard	Other tasks	Date	Deliverables
1	4/2/18	4/8/18					Form Team		
2	4/9/18	4/15/18	X				Form Team		
3	4/16/18	4/22/18	X	X			Public Material review	4/22/18	Goals
4	4/23/18	4/29/18	X	X			Public Material review		
5	4/30/18	5/6/18		X		X	Initial Findings preparation		
6	5/7/18	5/13/18		X	X	X	Initial Findings preparation	5/13/18	Initial Findings with Executive Summary
7	5/14/18	5/20/18		X	X	X			
8	5/21/18	5/27/18			X	X	Final report preparation		
9	5/28/18	6/3/18					Final and oral report preparation	5/29/18	Final Report with Executive Summary
10	6/4/18	6/10/18					Oral report preparation	TBD	Oral Report

Weekly status reports

We will keep your regularly updated with weekly status reports for weeks 3-10.

Initial Findings – 5/13/2018

On 5/13 we will provide a summary of our initial findings. This initial report will include:

- An Executive Summary
- An outline of the publicly available code and material review.
- A rundown of our Exploratory Data Analysis with notations on interesting data artifacts.
- An explanation of the tentative best model and modeling results that had been submitted to the Pump It Up data competition
- A proposed dashboard. This tentatively will include some map-based data, along with key measures to allow a quick insight into the state of the water system.

Final Report – 5/29/2018

On 5/29, we will include our completed final written report. In this report we will include:

- An Executive Summary
- The completed review of publicly available materials and code, along with an explanation of how we incorporated this information into our project
- A detailed explanation of the best model we found along with details about alternatives that were explored. We will include scoring information for all models and technologies considered.
- Recommendation for further investigation
- Detailed analysis of the expected costs and time associated with the upkeep of the modeling and dashboard systems

- In addition to a thorough review of our initial Exploratory Data Analysis findings, we will include any additional interesting aspects of the data that we discovered as part of the final modeling and dashboard creation
- The source code for our project
- The final dashboard

Oral Presentation – Week of 6/4

To wrap up this project we will present an oral report summarizing our findings that were included in the final written report. We expect this to last one hour. At the end of the presentation we will provide all materials presented.

6 Conclusion

DAFT PUMPS hopes to offer a more concrete and accurate analysis of the current water pumps in Tanzania as well as a more accurate prediction of the status of potential future water pumps in the area. Additionally, DAFT PUMPS will provide a consolidated view of the most up-to-date publicly available information regarding this project and the data set. This will allow us to build models more effectively by maximizing the value of our work. It will also position us to best speak to not just our process and model, but the distinctive value that we will provide.

The goals of DAFT PUMPS are clear: offer a consolidated analysis of the data and prior work, improve upon and offer higher accuracy levels regarding classification of water pumps, and a service-focused dashboard to show how we plan to utilize the model to direct our resources if we win the bid.

References

Darmatasia & Murni Arymurthy, A. (2016). *Predicting the Status of Water Pumps Using Data Mining Approach.* IWBIS 2016, 978-1-5090-3477-2/16.

Jimenez, A. & Perez-Fogue, A. (2011). *The relationship between technology and functionality of rural water points: evidence from Tanzania.* Water Science & Technology 63.5:2011.

Jimenez, A. & Perez-Fogue, A. (2011). *Water Point Mapping for the Analysis of Rural Water Supply Plans: Case Study from Tanzania.* Journal of Water Resources Planning and Management, ASCE September/October 2011: 441.

Welle, K. (2005). *Learning for advocacy and good practice*—WaterAid water point mapping: Rep. of findings based on country visits to Malawi and Tanzania. Retrieved from http://www.wateraid.org/documents/plugin_documents/waterpointmappingmalawitanzaniaweb.pdf