Predicting Water Well Functionality

Alexis Rouge Carrassat (amr439), Anthony Niznik (an533), Fan Liu (fl379)

Abstract
The country of Tanzania is currently facing a water supply challenge. Water wells are not fixed quickly, leading to thousands of people not having access to a clean water supply. The Tanzanian Ministry of Water is trying to determine whether water wells are functional, functional - in need of repair, or non-functional without having someone go to the well physically to check. The Ministry has requested the creation of a model that is able to determine which water wells are functional and which ones are in need of repair.

Contents

1. Exploratory Data Analysis

1.1 Data Characteristics

The data set comprises of a variable matrix with 39 features and 59,400 rows of water wells. The output we wish to predict is whether a water well is functional, non-functional, or functional-needs repair. The following are the number of columns (features) by type: 7 continuous, 27 categorical, 2 binary, 2 with all of the same value, and 1 date. Some continuous variables include *longitude*, *latitude*, *GPS height* (altitude of well), and *date_recorded* of the well. Please see Tables 1 and 2 to get a feel for the structure of the data set.

There were plenty of missing values and questionable zero values in columns. We had 3,056 missing values in the *permit* column and 3,622 missing in the *funder* column. Incorrect zero values occurred in the *longitude* column 1,812 times (Longitude Range: 29° E to 41° E) and 18,897 times in the *construction year* column (all other years >1959). To get a sense of where all the water wells are located, we plotted the water wells on a geographical map with blue indicating the well is functional, and red indicating non-functional (see Figure 1).

We see that wells form geographic clusters, especially around the coast and borders of Tanzania. In the next section, we discuss the insights obtained from visualizing the data.

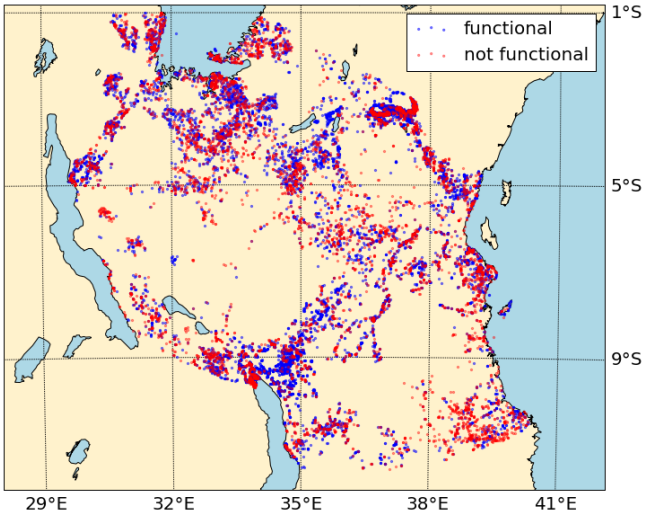

Figure 1. Locations of wells in Tanzania

1.2 Data Visualization

We first used histograms to visualize which variables could separate well functionality most distinctly. We found that whether a water well had a permit (legally authorized to exist) or not greatly affected the classification of the well. If a water well has a permit, it is much more likely to be functional than if it does not have a permit (see Figure 2 where a random sample of the data is plotted). This makes permit an important feature to include in our model.

Another question to address includes the following: how does the variability in continuous variables depend on the functionality of the well? We considered the variable *GPS Height*, which gives the height of the water well relative to sea level in meters (see Figure 3). We created a boxplot that displays the spread of the GPS Height for functional,

Table 1. Data set example

amount_tsh	date_recorded	gps_height	installer	. . .	longitude	latitude	status_group
6000.0	3/14/11	1390	Roman	. . .	34.938093	-9.856322	functional
0.0	3/16/13	1399	GRUMETI	. . .	34.698766	-9.856322	functional
25.0	2/23/13	576	Roman	. . .	37.460664	-9.856322	functional
0.0	1/28/13	263	UNICEF	. . .	38.486161	-11.155298	non-functional

Table 2. Summary statistics

Column Name	Mean	Standard Deviation	Minimum	Median	Maximum
longitude	35.98	2.56	29.61	36.65	40.3
latitude	-6.24	2.76	-11.65	-6.06	-1.04
gps_height	1002.37	618.08	-63.00	1154.00	2770.00

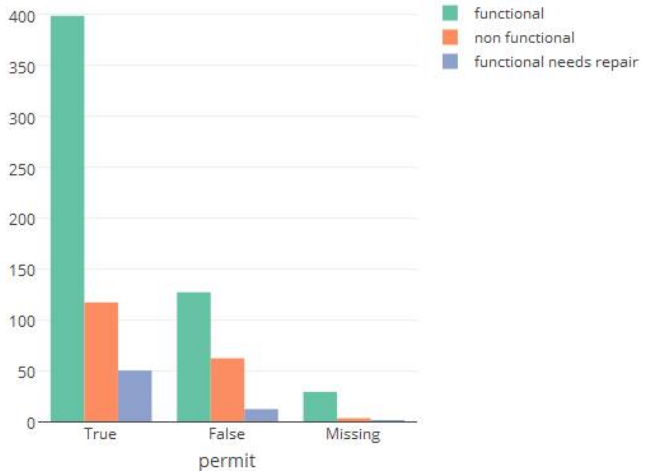

Figure 2. Number of wells based on permit

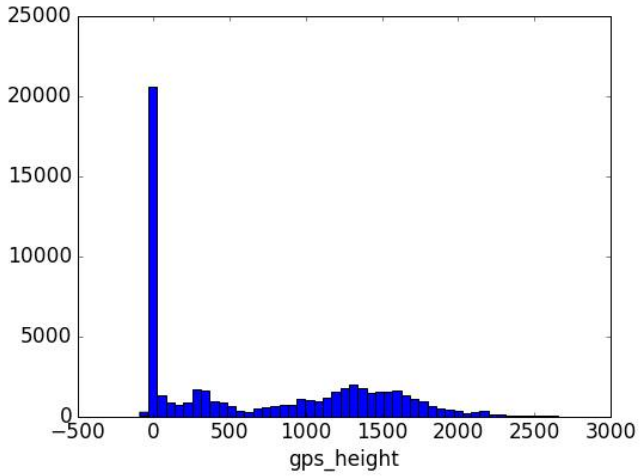

Figure 3. GPS height distribution (includes corrupt data)

non-functional, and functional-needs repair wells (see Figure 5). We see that the median GPS height of non-functional wells is slightly lower than the median of both functional and functional-needs repair wells. This implies that non-functional wells tend to have a lower GPS height than the other wells. Please note that the boxplot contains all values except for water wells at zero altitude because of our suspicion of corrupt entries (see Figure 4 and Section 1.4 for more details).

While plotting different features, we quickly realized that longitude and latitude were both important features for classification. Though we plotted the water wells onto the map of Tanzania, we wanted to get a better idea of where most of the functional wells may be located. We created a plot of longitude vs latitude this time to see if we can get a better sense of where the functional wells were located. We pinpointed a rectangular area in the center of Tanzania (shaded in blue) where over 70% of wells are functional (see Figure 6 where a random sample of the data is plotted). This result proved consistent when plotting with different samples of the data.

After this discovery, we considered the 21 regions in Tanzania and the percent of non-functional water wells in each region. When we created a bar plot that ordered the regions from lowest longitude (West) to highest longitude (East), left to right on the histogram, we found that the regions in the center of Tanzania (Iringa and Arusha) had the lowest percentages of non-functional wells (see Figure 7). This indicates that water wells in the center of Tanzania have a higher probability of being functional than wells farther west or east.

1.3 Cleaning the Data

1.3.1 Dates

We immediately left out of our model any columns that had the same value for more than 95% of their row values (2 features fit this description, *recorded by* and *number private*). Then, we changed the column that had calendar date of water well status check (day, month, year) to just the number of days from the current date to the earliest date of water well status check. So, if the earliest date any well was checked is 10/14/2002, then a well checked on 12/03/2013 would have 4,069 as its value. We also created two new binary columns, one column to indicate whether the water well was checked in the dry season and one column to indicate

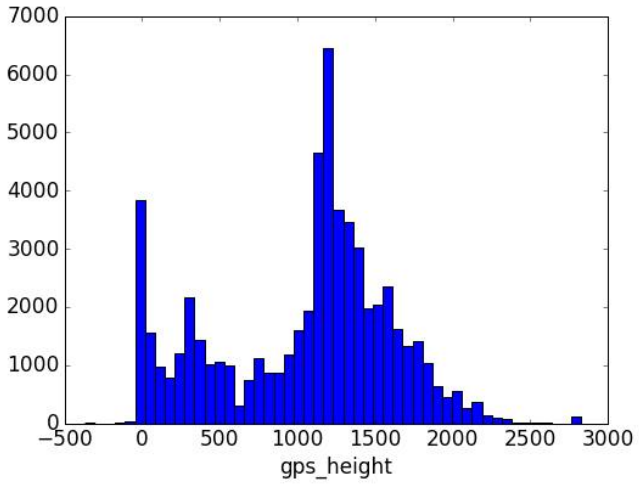

Figure 4. GPS height distribution (elevations corrected)

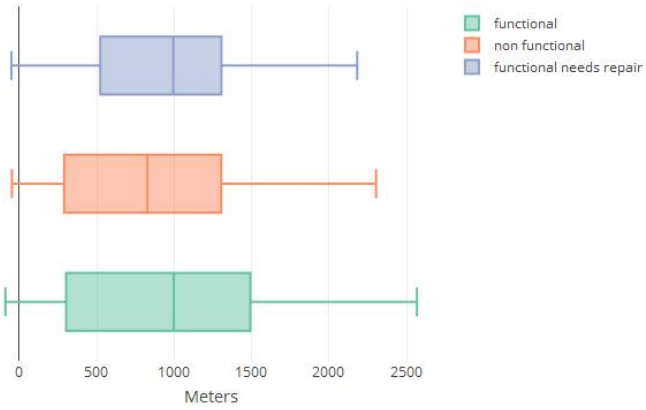

Figure 5. Well classification based on GPS height

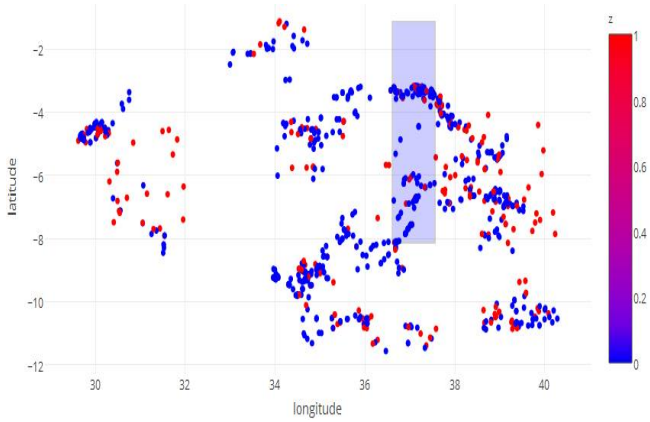

Figure 6. Longitude vs latitude of wells; functional wells (blue), non-functional wells (red)

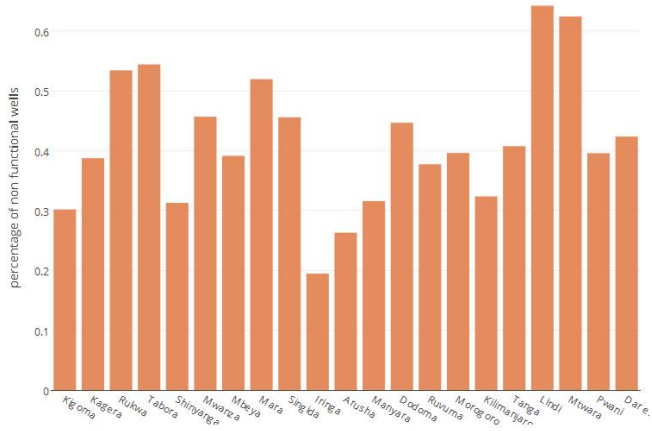

Figure 7. Barplot of percent of non-functional wells based on regions that are ordered by longitude (left to right)

whether the water well was checked in the wet season. The dry season include the months from June to October and the wet season includes months from November until May.

1.3.2 Many Unique Values

We also transformed the *installer* column (which company installed the well). Since there are 2,113 unique installer values and most rows belonged to the top two installers (See Table 3), we simply created three new binary variables (one for the installer with the most wells, one for the installer with the second most wells, all other installers placed in a *other* column, and a column for missing values discussed in the next paragraph).

1.3.3 One-Hot Encoding

For the categorical variables, we used one-hot encoding to change a variable of one column and d distinct values to d binary variable columns. This increased the number of categorical columns of our model to 226 columns. If values were missing from a particular column, we created a new binary variable that takes one if a value is missing and zero otherwise. In this way, we could retain some information, even for missing values and not have to delete an entire row just for 1 missing value. For logistic regression, we standardized continuous variables such as longitude and latitude of water well since regularized least squares is not scaling invariant. We standardized the data by subtracting the column mean of each column and then dividing by the standard deviation of each column for all continuous variables.

1.4 Corrupt Data

1.4.1 GPS Height

We noticed for the variable GPS height that the distribution of classification (functional, non-functional) is very different for the wells with zeros and the wells with a value not zero. We decided to plot the GPS height based on the longitude and latitude of each water well. We did this so that we could see where water wells are classified as having zero altitude to determine if they actually should be at zero altitude (see if they are corrupt). When we plotted the altitude of wells (see Figure 9), we found a

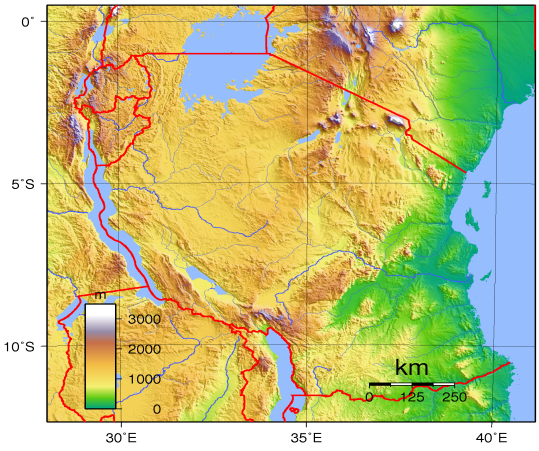

Figure 8. Elevation map of Tanzania (in meters)[1]

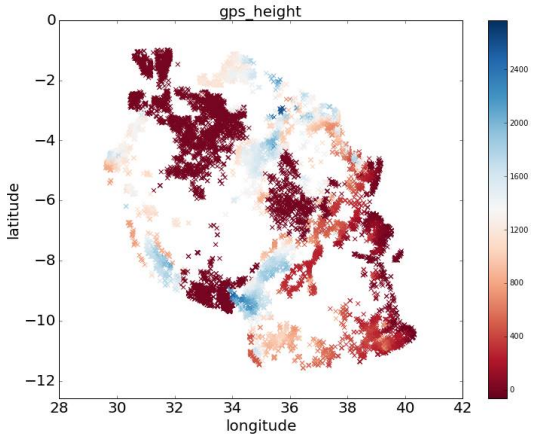

Figure 9. Geography of GPS height of well

Table 3. Category Frequency

Column Name	# Categories	Most Frequent	Frequency	# Missing
installer	2113	DWE	16255	3636
region	21	Iringa	5294	0
water_quality	8	soft	50818	0

big cluster of wells at zero altitude in the northwest part of Tanzania. When we compared this to the topography of Tanzania (see Figure 8), we found that the northwest region of Tanzania is about 1000 meters above sea level. This means that the water wells with GPS height of zero that are located in the northwestern part of Tanzania are incorrect and therefore corrupt.

1.4.2 Correcting Elevations

We verified using the Google Maps API that indeed most water wells that were assigned a GPS height of zero in the data set were at a higher elevation. To correct this corruption, we looked to substitute the incorrect zeros with the correct elevation using the Google Maps API. Since Google has a usage limit of 2,500 free requests per day[2] (and our data contains about 20,000 incorrect zeros), we decided to group the latitude and longitude into 1,600 squares. We did this by taking the range of latitude values that Tanzania spreads across and dividing that range by 40. We then did the same for longitude. With the 40 by 40 matrix, we were able to place 1,600 points on the area that encompasses Tanzania. With these points, we assigned each well to the point it should belong to by checking which point creates the minimum distance between the point and the well. With these assignments, we replaced the wrong elevation of wells with the new, approximate elevation assignment (See Figure 11 for more details).

1.4.3 Handling Zero Values in Population

For population, we decided to add a binary variable column that indicates whether the population is zero or greater than zero since zero appears very often (33% of the values are zero) and we do not have a means of verifying or substituting these values with more accurate ones. Additionally, the variance of population and GPS height are very high so we took the natural log of these variables to make sure that the values that are greater than zero but small will still have a significant impact on the model. This transformation is used for logistic regression and random forest which we will discuss in the next section.

2. Model Selection

2.1 Logistic Regression

For our initial model, we chose to use the logistic regression model (denoted by $h(x)$) because the water well problem is a classification problem. We first split the data into 70% training set, 15% validation set, and 15% testing set. We did this by using the randomsplit function. Randomsplit first generates a list of random indices and then it matches those indices to the actual values of the original array. Then it assigns the appropriate percentages of this random, matched list to each set. We used the validation set to calculate the λ that minimizes the percent of misclassifications. The objective function of logistic regression is defined as the following:

$$h(x) = log(1 + e^{(-yw^T x)}) + \lambda||w||_1 \quad (1)$$

Since we added many new variables with one-hot encoding, we chose l_1 for regularization because it encourages sparsity. Because logistic regression is a binary logistic model (1 if water well is a certain classification and 0 otherwise), we could not simply use one regression to perform the analysis. Instead, we created 3 different logistic regression models for each class

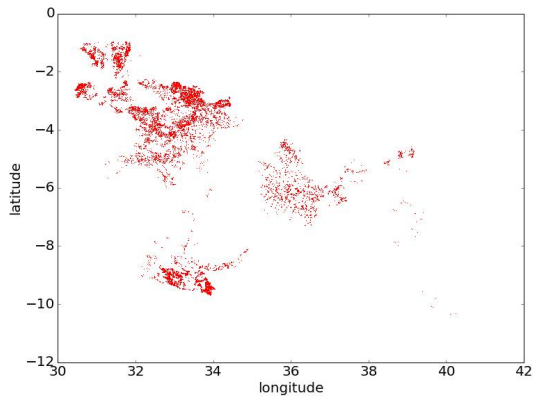

Figure 10. Locations of wells with zero GPS height

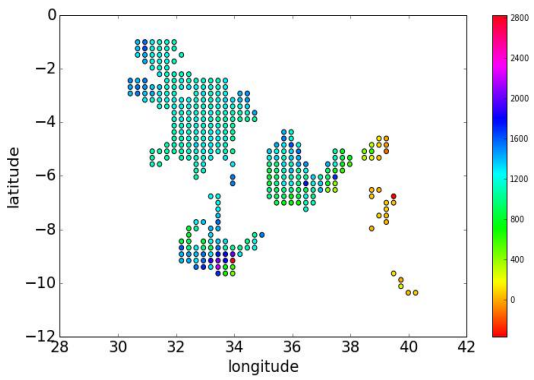

Figure 11. Points that assign new well elevations

i (one for each classification of functional, non-functional, and functional-needs repair). We then used that fact the logistic regression returns the probability that a water well is a certain classification to just simply take the maximum probability of the three models in order to classify the well (see Equation 2).

$$y = \underset{i}{\operatorname{argmax}}\, h_i(x) \tag{2}$$

2.2 Random Forest

Another model we considered is random forest. The way random forest works is that it creates groups of decisions trees (x) where each tree is assigned a certain number of features at random (while using the whole training set). Each decision tree is split on each feature based on maximizing the amount of information we can gain from each split (create the greatest reduction in entropy). When it comes time to predict, a majority vote is taken from the decision trees to classify each water well (see Figure 12). So, if a water well is classified by 10 trees, and 7 trees classify the well as functional, 2 trees classify as non-functional, and 1 classifies the well as functional-needs repair, then the well will be classified as the majority vote which is functional. For more information about random forest, please refer to chapter 15 in the textbook, *The elements of statistical learning: Data mining, inference, and prediction.*[3]

For our random forest model, we used 100 trees in our forest. We allowed trees to grow to a maximum depth of 30 and allowed trees to randomly split n times (here n is equal to the square root of the number of features chosen to split on). Random forest is not prone to underfitting since decision trees tend to overfit due to the fact that they can keep splitting to get a training error near zero. Also, random forest is not prone to overfitting since our maximum depth is a form of regularization that prevents each single tree from overfitting. Majority vote also helps prevent overfitting because it prevents overfit trees from weighing heavily on the classification.

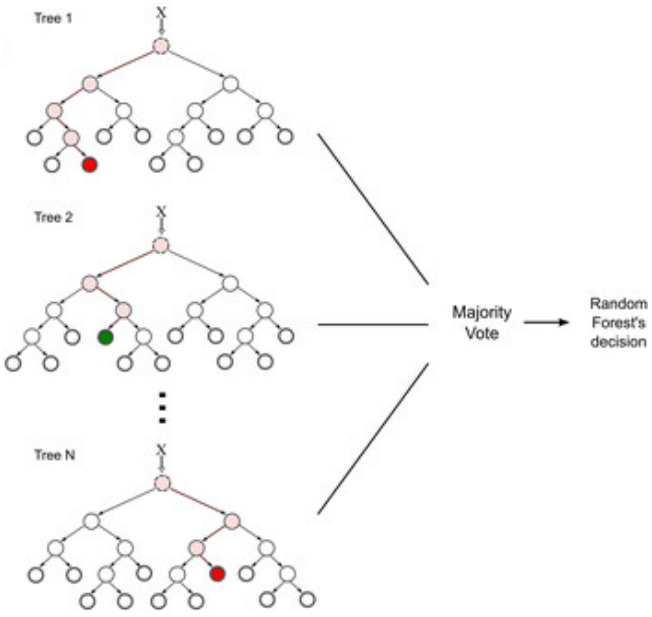

Figure 12. Random forest example[4]

3. Results

To test the effectiveness of both models, we calculated accuracy as the percentage of correctly classified water wells in our test set by the models. We found that logistic regression was able to classify with an accuracy of about 73%. On the other hand, random forest could classify with an astonishing 80% accuracy. A reason that the accuracy of the test data is low for logistic regression may lie in the fact that the training error for logistic regression is also low (about 74%) which implies that logistic regression has a relatively high bias.

When we replaced the incorrect elevations of water wells with more accurate elevations, we found that our accuracy increased a whole percentage to 81%! The significant impact of changing the incorrect zero values to a better approximation of elevation makes sense since the corrupted data affected over one third of the data set.

4. Conclusion

By using techniques we learned in class such as one-hot encoding, logistic regression, and regularization, we were able to classify the functionality of water wells in Tanzania fairly well with 73% accuracy. Since logistic regression is a linear model, it is not able to capture the inherit phenomena in the data set especially for continuous variables (such as latitude and longitude) because one can not necessarily create a line to separate functional and non-functional wells based on location. With random forest, we were able to achieve an accuracy of 81%.

We are fairly confident that our model will be able to generalize to new well data because random forest is very flexible and can adjust itself since it takes samples of the features to split upon. One thing to note is that we found longitude and latitude to be significant features in determining classification of water wells so if a new data set comes in without these features, our model may perform poorly. Also random forest is not as interpretable as logistic regression, so The Tanzanian Ministry of Water may prefer to use the logistic model since fewer variables can explain the classification of each water well. The Ministry will find our model quite useful because it has high accuracy and can handle corrupt or missing data. We hope this model can be a tool to bring clean water to a greater number of Tanzanians.

References

[1] Image obtained from: Wikipedia, "Geography of Tanzania"

[2] Google Maps Geocoding API Google Developers. (n.d.). Retrieved November 29, 2016, from developers.google.com/maps/documentation/geocoding

[3] Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer.

[4] Image obtained from: Machado, G., Mendoza, M. R., and Corbellini, L. G. (2015). *What variables are important in predicting bovine viral diarrhea virus? A random forest approach.* Vet Res Veterinary Research, 46(1). doi:10.1186/s13567-015-0219-7