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 being fixed quickly, leading to thousands of people not having access to a clean water supply.The Tanzanian Ministry of Water is trying to predict whether water wells are functional, functional - in need of repair, or not functional so that workers are allocated to fixing water wells rather than checking the status of wells. The Ministry has reached out to us to create a model that is able to predict which water wells are functional and which ones need to be fixed.

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 (Ordinal Values). There are 9 continuous variables, 26 categorical variables, 2 binary variables, 2 variables with all of the same values, and 1 date column. Some of the more interesting continuous variables include longitude and latitude, altitude of well (GPS height column), and construction year of the well. Some of the categorical variables include Region (like a U.S. State) and installer.

There were plenty of missing values and questionable zero values in columns. In particular, we had 3056 missing values in permit column, 3636 missing in installer, and 3622 missing in funder. Incorrect zero values occurred in the longitude column 1812 times (Longitude Range: 29° E to 41° E) and 18897 times in the construction year column (all other year in 2000s). Our approach to solve these issues will come addressed further in the data cleaning section. To get a sense of where all these water wells are located, we plotted the water wells on a geographical map with blue indicating the well is functioning, and red indicating not functioning (see figure 1).

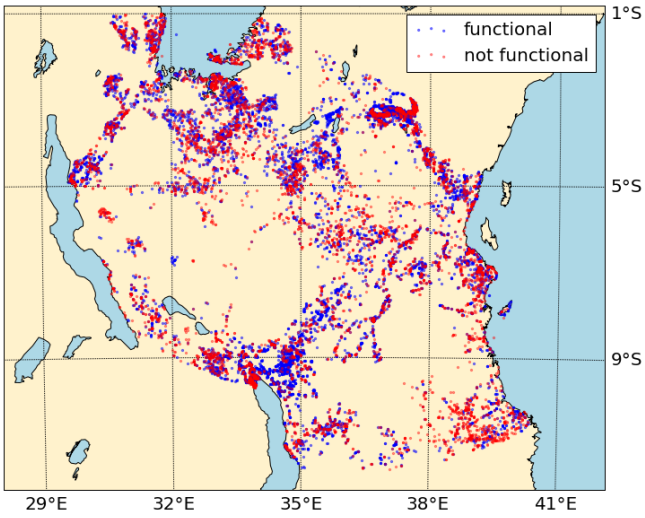

Figure 1. Locations of Wells in Tanzania

1.2 Data Visualization

We first used histograms to visualize which variables were able to 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 likley 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 a strong predicter of functionality of well for our model.

Another question that came up was how the variability in continuous variables could be different depending on the classifications. We looked into the variable called GPS Height that gives the height of the water well realtive to sea level in meters. We created a boxplot that displays the spread of the GPS Height for functional needs repair, non-functional, and functional wells (see figure 3). We see that the median GPS height of non-functional wells is significantly 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.

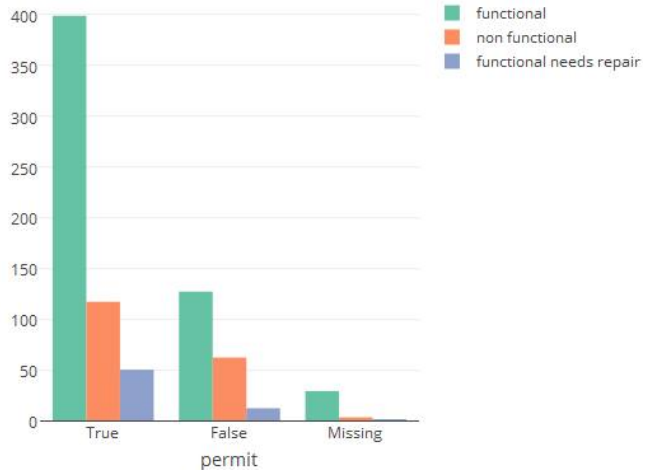

Figure 2. Number of Wells Based on Permit

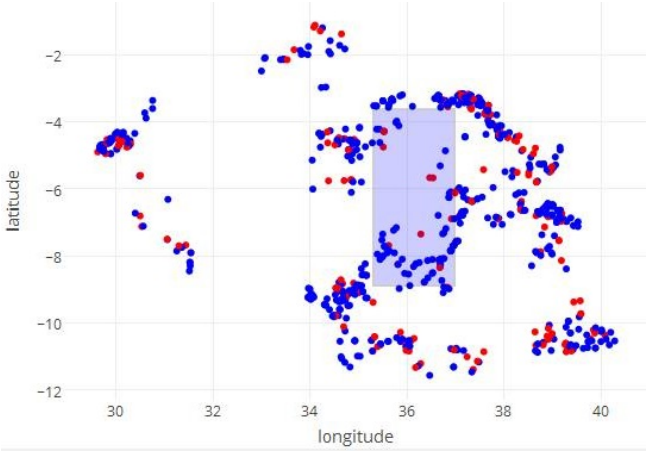

Figure 4. Longitude Vs Latitude of Wells; Functional Wells (Blue), Non-Functional Wells (Red)

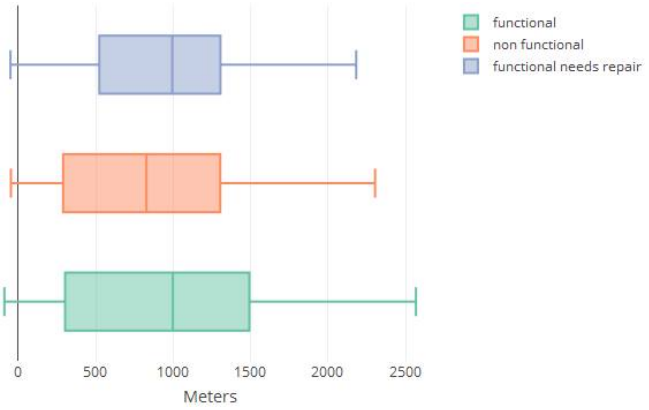

Figure 3. Well Classification Based on GPS Height

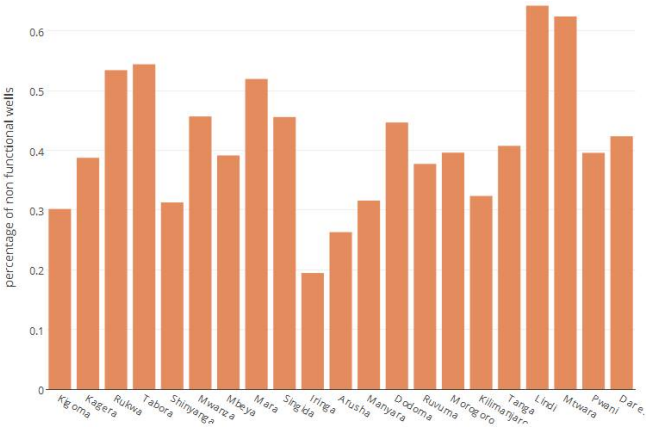

Figure 5. Barplot of Percent of Non-Functional Wells Based on Regions that are Ordered By Longitude (Left to Right)

While going through the different features, we quickly realized that longitude and latitude both were 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. We were able to pinpoint a rectangular area in the center of Tanzania (shaded in blue) where over 70% of wells are functional (see figure 4 where a random sample of the data is plotted).

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 to highest longitude (West to East), we found that the regions in the center of Tanzania (Iringa, Akusha, and Manyara) had some of the lowest percentages of non-functional wells (see figure 5). This means that we can explore these regions further to see which characteristics they may have that leads to a high percentage of functional wells.

1.3 Cleaning the Data

We immediately left out of our model any columns that were nearly duplicates of each other (more than 95% of same row values amd measuring same thing). Then, we changed the column that had calendar date of well status check (day,month,year) to just the number of days from the current date to the earliest date of 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 4069 as its value. We also transformed the installer column (which company installed the well). Since there are 2113 unique installer values and the majority of rows belonged to the top two installers (outliers compared to the rest of the data), 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.)

For the categorical variables, we used one-hot encoding in order to change a variable of one column and d distinct values

to d binary variable columns. This increased the number of columns of our model from 28 to 220 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 noticed for some variables such as population and GPS height that the distribution of classification (functional, not functional) is very different for the wells with zeros and the wells with a value higher than zero in our data visualization (see figure 3). Due to this, we added a binary variable column that indicates for population whether it is zero or greater than zero since zero appears very often (33%). Additionally, the variance of population and GPS height in well is very high so we took the natural log of these variables to make sure that values that are greater than zero but small will still have a significant impact on the model.

2. Model Selection

2.1 Logistic Regression

For our intial model, we chose to use logistic regression model (denoted by $h(x)$) because the water well problem is a classification problem. We first split the data into 60% training set, 20% validation set, and 20% testing set. We used the validation set in order to calculate the λ that minimizes the percent of misclassifcations. 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 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 perfrom the analysis. Instead, we created 3 different logistic regression models for each class i (one for each classification of functional, non-functional, functional - needs repair). We then used that fact the logistic regression returns the probability that a water well is a certain classifcation to just simply take the maximum probability of the three models in order to classify the well (see equation 2).

$$y = \underset{i}{\text{argmax}}\, h_i(x) \quad (2)$$

2.2 Random Forest

Another model we considered is random forest. The way random forest works is that it takes a collection of random samples from the training set while maintaining the underling distribution. Then, it creates groups of decisions trees (x) where each tree is assigned a certain number of features at random. 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 each group of decision trees to classify each water well (see figure 6).

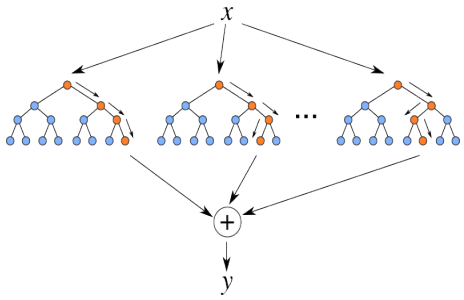

Figure 6. Random Forest Example

For our random forest model, we used 30 trees in our forest. We allowed trees to grow to a maximum depth of 20 and assigned trees to randomly pick the squart root of the total number of features to split upon. The number of features to split upon and the maximum depth are parameters that we will fine tune later in the project. Random forest prevents underfitting since trees are randomly assigned to a subset of features to split on which helps to account for all the different characteristics water wells may have (flexible). Also, random forest is not prone to overfitting since the majority vote of a group of trees prevents overfit trees from weighing heavily on the classification.

3. Results

To test the effectiveness of both models, we calculated accuracy as the percentage of correctly classified water wells in our testing 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 was able to classify with an astonishing 80% accuracy. An indication 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% also) which implies that logistic regression has a relatively high bias.

4. Next Steps

To improve our model, we will look to transform some of our variables in order to capture information better than simply including data in its raw state. For example, we can add a binary column that indicates whether the well was checked in the dry season or the rainy season . This can show trends of how weather seasonality may relate to the probability of a water well being functional. Our ultimate goal is to be able to classify water wells with the most accuracy and even to classify better than the current best model for this problem (classifies with 82% accuracy). Another route worth considering is to help the Tanzanian Ministry of Water learn about which features lead to the highest liklihood of a water well functioning properly. This could help the Ministry to know how to build new water wells so that they are more likely to work.