Using Installation and Management Data to Predict Waterpoint Failure in Tanzania

Final report prepared for the Coursera Data Analysis and Interpretation Specialization

June 12, 2016

INTRODUCTION

The purpose of this project is to identify the best predictors of water pump (a.k.a. waterpoint) operational status in the country of Tanzania based upon several factors related to its installation and operation. Such factors include the type of pump, when it was installed, its location, the water source/quality, and who manages the operation of the waterpoint.

Being able to predict which waterpoints will fail ahead of time, or even just which ones may have already failed or need repair that have not been reported, can improve the scheduling and cost-effectiveness of maintenance operations by the Tanzanian Ministry of Water (TMW). Ultimately, the goal is to have clean, potable water always available to all Tanzanian communities.

METHODS

Sample

The data used in the analysis came from DrivenData Inc. as part of a hosted data analysis competition named “Pump it Up: Data Mining the Water Table.” The competition data was provided by the TMW from its Taarifa waterpoints dashboard. For details about the competition, see the DrivenData web site at https://www.drivendata.org/competitions/7/.

The competition data consists of 1) a competition training data set of N=59,400 data records (the competition training data); 2) a competition test data set of N=14,850 data records (the competition test data). Each data record pertains to a single waterpoint maintained by the TMW. For each waterpoint, the data describes its location, installation, usage, management, water characteristics, and waterpoint type. Additionally, the training data set identifies the functional status of each waterpoint whereas the test data set does not include this data.

The competition information does not state specifically how the training and test data sets were created. This analysis assumes that they a random split from a single complete data set or were somehow separately randomly selected from a larger volume of data. It is further assumed that the two data sets are entirely disjoint, meaning that no waterpoints exist in both data sets.

Measures

The outcome variable for the analysis is waterpoint status, which is a categorical variable with possible values of “functional” “functional needs repair”, and “non functional”.

Five categories of predictors were chosen to be evaluated for use in the prediction model, providing a total of 15 individual candidate predictor variables. The candidate predictors are:

Waterpoint location

- Variables: region id, district id, ward name, village name

Waterpoint installation

- Variables: installing organization name, funding organization name, permitted installation (yes/no), construction year

Waterpoint management

- Variable: type of managing organization (private operator, water user group, etc.)

Waterpoint type

- Variables: pump type (hand pump, standpipe, etc.), water extraction type (gravity, submersible pump, etc.), output quantity (enough, insufficient, etc.)

Water source

- Variables: geographic water basin name, water source type (river, spring, etc.), water quality (salty, soft, etc.)

Other variables in the competition data set were not chosen for the analysis because the data was 1) redundant (e.g., region names were redundant with region ids); 2) financial related (e.g., how water payments are made); or 3) not otherwise related to the physical nature nor use of the waterpoint (e.g., GPS location).

Analyses

For this analysis, the competition training data set was randomly divided into a working training data set (hereafter referred to as the training data set) containing 70% of the competition training data and a validation data set containing the remaining 30%. The validation data set was used to test model accuracy prior to applying the model to the competition test data set; which was done only once to get the final predictions.

All of the chosen candidate predictors are categorical variables with the exception of waterpoint construction year, region id, and district id. Even so, only construction year has any numerically

useful properties. Region id and district id are just numeric equivalents of the region and district names. Thus, no standardization of any of the variables was done. However, all of the non-numeric categorical variables were converted to numerical equivalents to suit the model building method applied (see below).

The only missing data noted were zero values for construction year in some observations. A zero construction year value occurred in too large a percentage of the data (34.8%) for those observations to be filtered out. It was decided to leave the zeros in place as an indicator of unknown construction date.

A random forest technique was used to fit prediction models with the candidate predictor variables to derive waterpoint status. This technique was chosen because it is recognized as a highly effective means for building prediction models with categorical variables. The forest size was limited to 100 trees because review of the error rate versus tree count showed 100 trees to be more than sufficient to achieve a low error rate.

The predictive accuracy of any model was measured as the percentage of correct waterpoint status predictions. During the model building and selection phase of the analysis, the estimated out-of-bag (OOB) error rate was used. The accuracy of the final prediction model was determined by applying it to the validation data set.

RESULTS

Descriptive Statistics

There were 59,400 unique waterpoints identified in the competition training data set. After accounting for missing construction year data (35% of the observations in the data set), the waterpoint construction years are in the range 1960-2013. Waterpoint operational status was broken down as follows: 54% were functional, 7% were functional needing repair, and 39% were non-functional. The randomly created training and validation data sets for model creation had similar breakdowns.

Summary statistics for the candidate predictor variables are not meaningful because of their categorical. Most have a large number of possible values, making bar charts or box plots impractical.

Bivariate Analysis

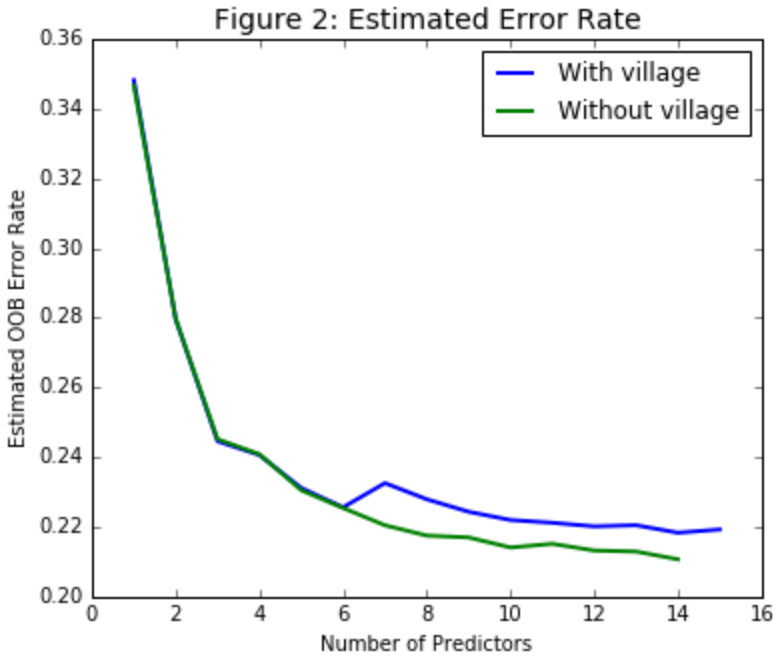
Each candidate predictor variable was singularly used to fit a random forest model as a first pass test to eliminate the variables with the least impact on predicting operation status (based on estimated OOB accuracy). Figure 1 shows a table of the results. The OOB accuracy of each candidate predictor alone was in the 54%-65% range.

Figure 1: Accuracy Rates for Single-predictor Models

Predictor	Estimated OOB Accuracy
ward	0.651900
output quantity	0.647860
water extraction type	0.623569
pump type	0.617412
installer	0.612266
funder	0.598124
village	0.585402
construction year	0.578018
region code	0.567917
water quality	0.567220
water basin	0.560029
district	0.556830
water source type	0.544901
type of managing organization	0.544661
permitted installation	0.542593

Multivariable Analysis

Next, a series of models were fit starting with the most accurate candidate predictor variable and adding additional variables one-at-a-time in descending order of their individual accuracy (see Figure 1). The resulting error rates are shown in Figure 2.



The blue line shows the first pass of the analysis. There is a clear increase in the error rate when the seventh variable (village) is added. A second pass at the analysis was made leaving the village variable out of the set of predictors. The green line shows the second pass results. Clearly, village is a confounding variable in the data even though individually it has a 58% prediction accuracy.

Figure 2 clearly shows error rate mostly leveling off after eight predictor variables. There is some minor further reduction in error rate as more predictors are added, but that may be a case of over fitting the training data set rather than a useful increase in predictive capability. The eight-predictor model had an estimated OOB accuracy of 78%.

The top eight predictors – ward, output quantity, water extraction type, pump type, installer, funder, and construction year - were refit to the training data set to create for the final model. This model was then used to make waterpoint operational status predictions using the validation data set. Figure 3 shows the resulting confusion matrix.

Figure 3: Validation Set Prediction Confusion Matrix

True \ Predicted	Functional	Functional Needs Repair	Non-functional	All
Functional	8418	294	986	9698
Functional Needs Repair	650	412	237	1299
Non-functional	1532	137	5154	6823
All	10600	843	6377	17820

The overall accuracy is 78%. This compares very favorably with the 78% estimated OOB accuracy from the model using the training data set.

The model performs reasonably well at predicting “functional” and “non-functional” operational status, but is much less accurate regarding “functional needs repair” status.

CONCLUSIONS/LIMITATIONS

This project used random forest analyses to identify the best predictors of water pump (a.k.a. waterpoint) operational status for N=59,400 waterpoints in the country of Tanzania based upon several factors related to its installation and operation. This data was acquired from DrivenData Inc. as part of a hosted data analysis competition named “Pump it Up: Data Mining the Water Table.” as provided by the Tanzanian Ministry of Water.

After reviewing the estimate accuracy of several models, 8 of the 15 candidate predictor variable were chosen to create the final prediction model. These 8 predictors resulted in a 78% overall prediction accuracy when the model was applied to a validation data set. This exactly matched the 78% estimated OOB accuracy when the model was created. The ward name variable accounted for 65% of the prediction accuracy, with output quantity and water extraction type contributing an additional 9%.

Even though the model achieved a respectable overall 78% accuracy, it was not uniform. The accuracy of predicting functional waterpoints was 87%, 76% for non-functional waterpoints, and only 32% for functional waterpoints needing repair. It is not clear why the “functional needs repair” status was more difficult to predict than the other two. Perhaps the reason is because it is a special case of the functional status and there is nothing in the data to be able to distinguish between the two.

The model is only capable of reactive prediction; meaning predicting the next likely waterpoint failures given the occurrences of a failures in a nearby or similar waterpoint. The primary limitation of the model is that the predicted results will be the same every time it is run until a real change occurs in the operational status of a waterpoint; at which time the model would need to be rebuilt – not just re-run – to see if the predictions for other waterpoints change.

Ideally, one might want to run the model weekly or monthly and get a new set of predictions based upon time passing and the probability of waterpoints failing possibly changing as a result. Building a model for this would require including data about past related events. How long has a waterpoint been operational or when was its last failure? How often has it failed in the past? What is its mean time between failures? Why did it fail? And, so on. Future efforts to expand the usefulness of the model should consider taking these into account. If not readily available then obtaining, saving, and processing snapshots of the TWM data at regular time intervals (e.g., daily or weekly) would be a start to developing a more pro-active prediction model.