2020

PREDICTIVE MODELS FOR MAINTENANCE OF WATER PUMPS IN REMOTE LOCATIONS

FINAL REPORT PREPARED FOR THE DATA ANALYSIS AND INTERPRETATION SPECIALIZATION

MO VENOUZIOU - MVENOUZIOU@GMAIL.COM

INTRODUCTION

Clean, potable water is a basic human need. Beyond mere *access* to water, proximity to a reliable water source is a major factor in determining the educational and economic opportunities available to adjacent communities. Because of this, the development and maintenance of water sources in remote and/or impoverished communities is a major area of focus for many government and nonprofit humanitarian organizations. (Human Rights to Water and Sanitation, 2020)

This paper develops a predictive model for determining the maintenance needs of water pumps in Tanzania. Specifically, this is a supervised learning problem to classify pumps as functional, in need of repair or completely inoperable. The model achieves 79.63% predictive accuracy on the DrivenData test set using decision tree models.

It is the hope that these results will 1) lead to more effective allocation of resources in Tanzania and other countries facing similar challenges and 2) provide a starting point for analyzing other remote infrastructure needs, such as internet and phone networks.

METHODS

Sample

This analysis is based on a DrivenData.org repository of data sourced from the Taarifa Platform and the Tanzanian Ministry of Water. (Pump it Up: Data Mining the Water Table, 2020)

I narrowed down the labelled dataset to the 59,369 test results taken between 2011 – 2013, representing 57,489 individual water pumps across Tanzania. According to data from the Tanzanian Ministry of Water there are a total of 87,436 monitored pumps throughout the country (Tanzanian Ministry of Water, 2020).

DrivenData.org also provided a test set of unlabeled data on 14,850 pumps. All test scores reported in this paper reflect official scores generated by DrivenData.org on this test set. Any training or validation scores reported are self-generated from the labelled training data.

It is unclear how the specific sample locations were chosen, or why some pumps appear more than once in the set. There was no labeled or unlabeled data on 15,097 pumps tracked by the Tanzanian Ministry of Water. It is likely that the missing data is a randomly chosen subset held back for use in other DrivenData competitions.

Measures

Each sample included 44 predictive features for an individual water pump, plus a target label classifying the pump as ‘”functional,” “function but needs repair” or “non-functional.” The provided features relate to pump construction/ funding, environmental factors and ongoing maintenance.

This analysis narrowed down to the following 11 “core” features found to be important for analyzing the dataset, plus 13 “optional” features that offer a small improvement in prediction accuracy.

Core Features

Age of Pump (Log-scaled)

Extraction Type Used

Geographic Water Basin

Management Group

Proper Permitting Obtained (Yes / No)

Public Meeting Held (Yes / No)

Water Point Type

Water Quantity

Water Quality

Water Source

Water Usage Pricing Structure

Additional Features (optional)

Age of Pump (unscaled)

Funder

GPS Height of Pump

Installer

Management Organization

Population Size Served

Region

Sub-village

Total static head (water pressure)

Type of Water Source

Type of Pump (above-ground component)

Type of Pump (below-ground component)

Ward

Analysis

I began by separating the labelled data into training and validation sets using an 80% / 20% split. Missing quantitative data was filled in using the training set mean, while missing categorical data was classified as “other.” To prevent overfitting of features attaining 10 or more possible values, such variables were binned into 6 subgroups using equally spaced percentiles.

Next, I performed logistic regression and examined frequency graphs on the training set to gain an intuition on what features will have predictive power. Examining p-values and the size of coefficients and removing redundant variables allowed us to narrow down features to those listed in the “Methods” section.

I then ran the data through various decision tree and neural networks and chose the most successful model on the validation set.

There is a significant quantity of missing values in the data. This problem was especially pronounced in the important “age” feature, which was missing in over 30% of the observations. I ran models using four variations: 1) filling missing data with training means, 2) filling with medians, 3) dropping missing age data and filling the rest and 4) dropping all missing data.

Filling all data with means proved to be the superior approach. Dropping data led to severe overfitting that could not be overcome, resulting in much poorer test scores. Using medians was

only marginally worse than means. The below discussion describes the case where missing data is filled in, unless otherwise noted.

Baseline for Comparisons

At the very low end, the majority class model is the baseline for comparison. This model predicts all pumps to be functional, with an anticipated 54.3% accuracy based on the frequency distribution in the dataset. At the high end, the top performing model on DrivenData.org's public leaderboard achieves 82.94% test accuracy (as of 11/29/2020).

RESULTS

Descriptive Statistics

Examining a logistic regression with quantitative and binary features I was surprised to find that, while all have low p-values, the features "population", "water pressure" and "GPS height" have very small coefficients. Perhaps the large number of missing data (filled with mean value) interfere with their influence.

"Age" should be one of the most predictive factors involved with pump status, therefore I generated this variable along with its logarithm, log(age + 1), for use in the models. More than 30% of age-related data is missing, however, and much of its predictive power has been diminished.

Figure 1: Logistic regression for Quantitative and Binary Features vs Pump Status.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.2774	0.057	4.864	0.000	0.166	0.389
age	-0.0364	0.002	-15.200	0.000	-0.041	-0.032
Q("age (log scaled)")	-0.1794	0.032	-5.548	0.000	-0.243	-0.116
population	7.571e-05	2.17e-05	3.483	0.000	3.31e-05	0.000
water_pressure_amount_tsh	1.717e-05	5.48e-06	3.132	0.002	6.43e-06	2.79e-05
gps_height	0.0006	1.91e-05	28.815	0.000	0.001	0.001
Q("permit (y/n)")	0.1335	0.020	6.562	0.000	0.094	0.173
Q("public_meeting (y/n)")	0.4082	0.032	12.599	0.000	0.345	0.472

A second logistic regression, now with the categorical variables, shows all of them to be influential. Water Source, water pricing, extraction type, water quantity, water quality, water point type and basin also exhibit large variability in frequency bar graphs.

Only two features available to us relate to ongoing maintenance: water pricing and maintenance organization. Because of this special status, both were included as core model features.

Predictive Models

I modeled this classification problem using various neural network and decision tree architectures. In all cases the decision trees matched or outperformed the neural network.

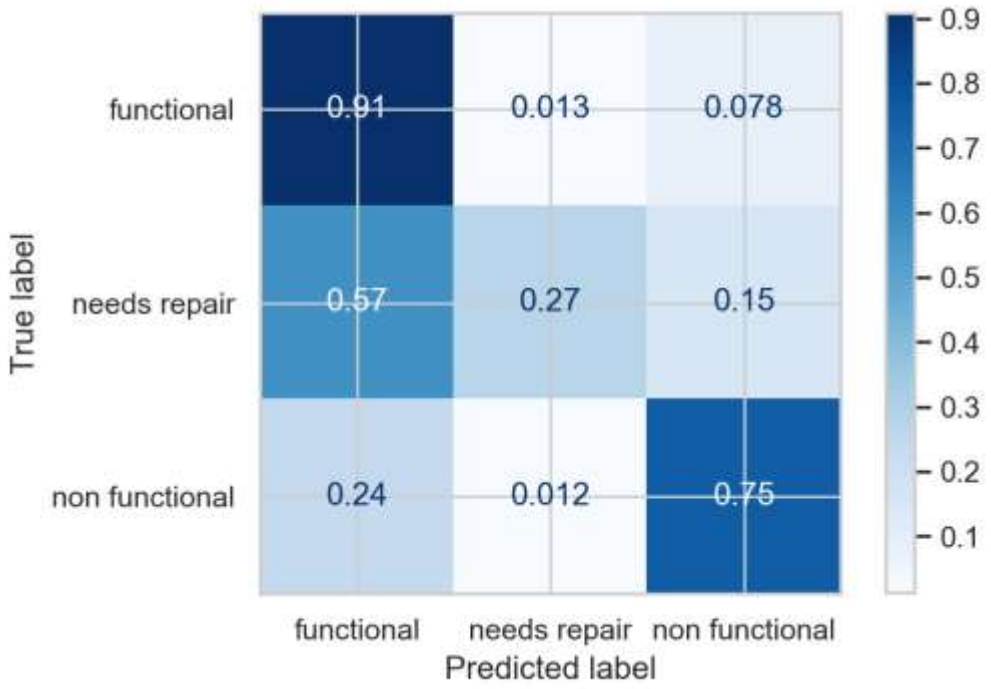
Using *only the core features* identified in previous sections, together with scikit-learn's Random Forest Classifier (with n=200 estimators and minimum sample split = 25) resulted in a test score of 77.3% and 3.4 percentage point test / train variance.

This result can be improved marginally by adding in the "optional" features. None seem to harm the model, so the final model included all of them. Using scikit-learn's Extra Trees Classifier (with n = 100 estimators, max_features = 'auto' and a minimum sample split of 50 to control overfitting), achieved a 79.63% test score with 2.6 percentage point train / test variance. This is

over 25 percentage points better than the baseline majority class model within 3 percentage points of the top comparison model.

Most of the models' error are due to false positives from over-predicting pumps to be functional. It is successful at predicting functional and non-functional pumps, but struggles with pumps in need of repair. Weighting classification labels could reduce this effect, but at the expense of overall predictive accuracy. The ultimate choice of whether to make this tradeoff requires information about the costs associated with the different types of error.

Figure 2: Confusion Matrix representative of the models' prediction results.



Conclusions

The ten "core features" provide nearly all the predictive power found in each model, bringing them to within 2.33% points of the best test score achieved. Adding in all features resulted in a 79.63% test score with 2.6 percentage point train / test variance.

Core Features

Installation:

Extraction Type Used, Water Source, Proper Permitting Obtained (Yes / No), Public Meeting Held (Yes / No), Water Point Type, Geographic Water Basin, Water Quality

Ongoing Maintenance:

If / How Community is Charged for Water

Management Group

Factors outside our control:

Age of Pump (Log-scaled)

Water Quantity

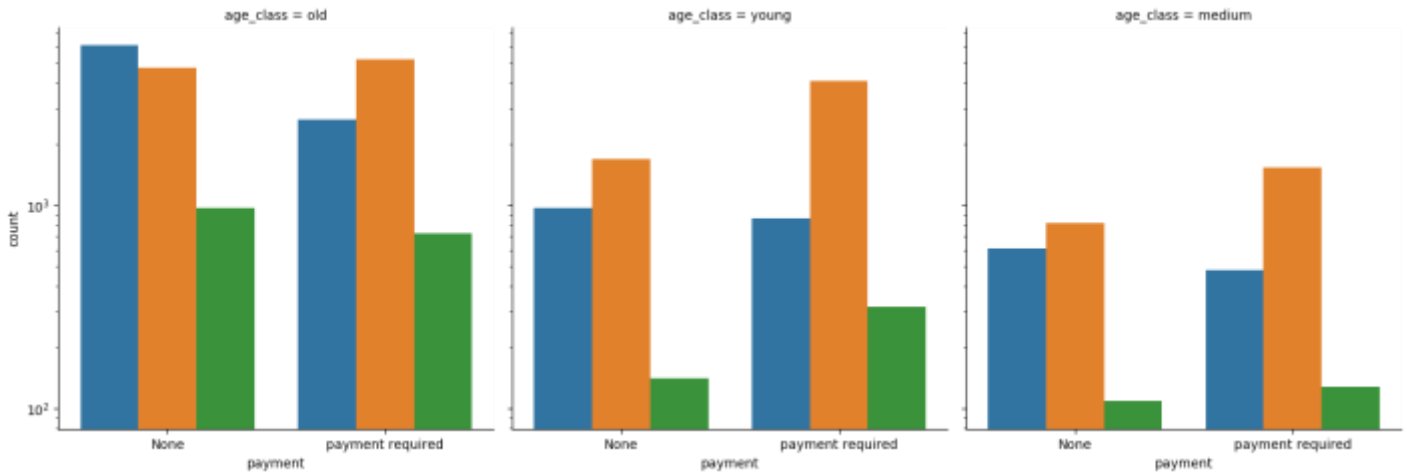
There are two main roles for funders and policymakers: increasing access and properly maintaining the water infrastructure. The core features above may help policy makers focus on relevant predictive factors within their control.

I find the two “ongoing maintenance” features of greatest interest. After a pump has been built, these are the only model variables under policy-makers control.

Requiring people to pay for water use appears highly associated with a pump’s operating status, even after controlling for age. Policy makers may consider implementing some payment structure, albeit at a low enough rate so as not to burden the populace.

Management structure is also associated with pump operating status, even after controlling for age. While not as strong an effect as water payment, it has predictive power among functional pumps.

Figure 3: Pump status counts by payment plan and age group. Pumps requiring payment show better functioning status



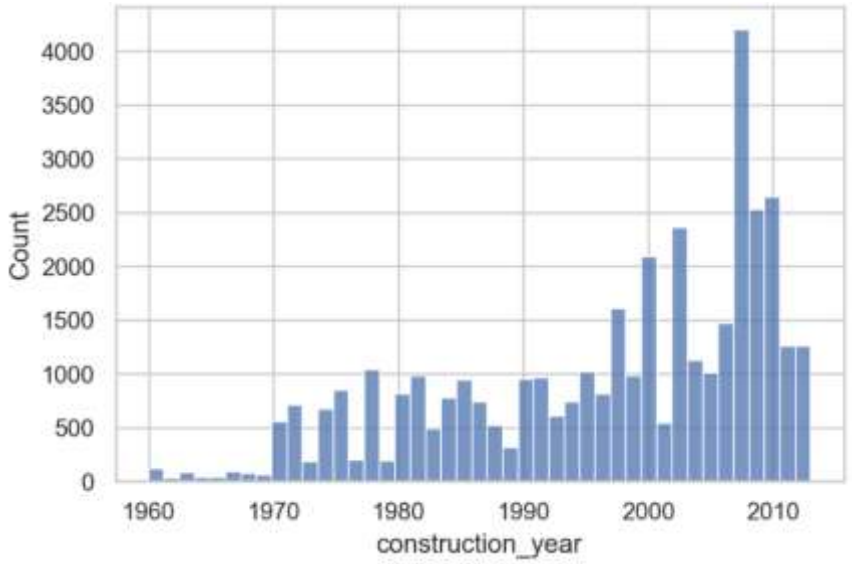
Limitations & Future Research

While the models have high predictive power, they suffer from high false positive rates incorrectly predicting pumps to be functional. Should they be reweighted to trade off accuracy for lower false positives?

What excess repair would be caused from under-maintaining pumps? Is there redundancy in the system, or would a pump being in disrepair cause a community to lose access to an acceptable water source? Further research is needed to better understand all the costs involved.

Another challenge arises from confounding factors related to changes in Tanzania's political environment. Some pumps under consideration were constructed as early as the 1960's, around the time the nation was formed through emancipation from the British colonial system and the merger of Tanganyika and Zanzibar.

Figure 4: Distribution of pump construction year. Note: construction year is missing from over 30% of observations



Since that time there have been major fluctuations in political structures, international aid and economic challenges (Britannica, 2020) which may have had an effect on maintenance and construction of basic infrastructure such as water supply.

A more detailed study of these effects is warranted, perhaps through comparative studies of other African nations with similar economic conditions but different politics.

Works Cited

Britannica, E. (2020, December 6). *Tanzania: Challenges into the 21st century*. Retrieved from Encyclopædia Britannica: https://www.britannica.com/place/Tanzania/Challenges-into-the-21st-century

Human Rights to Water and Sanitation. (2020, 11 13). Retrieved from United Nations UN-Water: https://www.unwater.org/water-facts/human-rights/

Pump it Up: Data Mining the Water Table. (2020, 11 13). Retrieved from DrivenData: https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/

Tanzanian Ministry of Water. (2020, November 30). Retrieved from Water Point Mapping System (WPMS) Tanzania: http://wpm.maji.go.tz