PREDICTING WATER PUMP FUNCTIONING TO AID TANZANIAN COMMUNITIES

Report by Kishore P. V.

1. INTRODUCTION TO THE RESEARCH QUESTION

This study is aimed at identifying the relationship between functioning of water pumps (response variable) and following set of features- amount water available to waterpoint, altitude of the well, location, population around the well, year of construction, extraction type used by waterpoint, management of waterpoint, payment, quality, source of water and kind of waterpoint. The end goal is to accurately predict the functioning condition of pump, given the features (explanatory variables).

As an engineer, I explore ways to use my knowledge for the welfare of people around me. Learning that predicting the condition of water pumps can help ensure access to clean and portable water to communities across Tanzania, I am motivated to work on this project.

Accurate prediction of failure of water pumps can help maintenance operations be carried out in a timely manner. This can ensure undisrupted availability of clean and portable water to lots of communities in Tanzania.

2. METHODS

2.1 Sample

The dataset used for this study is available from DRIVENDATA, as part of their "Pump it Up: Data Mining the Water Table" competition. This dataset is derived from data made available by Taarifa (www.taarifa.org) and the Tanzanian Ministry of Water (www.maji.go.tz).

The sample consists of 59400 entries, each enumerating the functional condition of a water pump along with details of 40 waterpoint characteristics like co-ordinates of pump location, geographic region where pump is located, source of water, age of water pump etc. As data was available as part of the competition, no sampling criterion was used to obtain data from the two above mentioned sources.

2.2 Measures

Following are the features included for analysis:

1. Amount of water available to the waterpoint - The variable has a lot of values=0 (~40000). These values are retained as 0, which signifies unknown value.

2. Altitude of the well - The altitude of some wells are wrongly specified as 0 or negative values. These values are recoded as 0.

3. Organization that installed the well - Entries with value "Not known" or "0″ or NAN are recoded as "other".

4. Longitude and latitude of the pump location- Latitudes with value -2.000000e-08 and longitudes with value 0.0 are invalid locations.

5. Geographic water basin- This is a categorical feature that takes 9 different values,.

6. Geographic location- It is captured in the lga, region codes and district codes variables. The string values (of lga feature) are converted to lower case7. Population around the well (population) - Invalid or unknown values for population are mentioned as 0.

8. Who operates the waterpoint (scheme_management, scheme_name) - Both are categorical features.

9. Year of pump construction- Unknown values are coded as 0.

10. The kind of extraction the waterpoint uses- This feature has 7 categories.

11. How the water pump is managed- It is a categorical variable with 12 categories.

12. What the water costs – It is categorical variable whose values range from "no payment" to "pay per bucket".

13. Quality of water- It is a categorical feature.

14. Quantity of water available- It is a quantitative feature.

15. Source of water- It is a categorical variable with 10 categories

16. Kind of waterpoint- It is a categorical variable with 7 categories.

The status of pump is the response variable. It can take the values- functional, functional but requires repairs or non-functional. Categories of categorical variables are recoded as numbers.

Invalid values are not dropped, but are rather recoded in a consistent manner. This is done because, in the situation where unknown values appear in the test case, the model should be able to provide a good prediction. Also, removing the invalid values significantly reduces the size of the training dataset, and in the process would be losing a lot of information.

3. ANALYSIS

To get a preliminary idea about the features, summary statistics of each of the features were analysed. Mean, standard deviation, min/max and four quartiles of quantitative features, count and most common category for categorical variables were calculated. Univariate histogram (for quantitative features) and bar graphs (for categorical features) were constructed.

Bi-variate bar graphs were constructed to analyse how functioning of pump varies with change in each of the features. The quantitative variables are converted to categorical (cut into four categories on 25%, 50% and 75%ile boundaries). The *status of pump* (response variable) is remapped as: not functional to 0, functional or function but needs repairs to 1.

The response variable is categorical and the explanatory variables (features) are categorical or quantitative. For testing association between categorical features and status of water pumps (response variable), Chi-square test was conducted. The quantitative features were converted to categorical and Chi-square test was performed to analyse the significance of their relation with status of water pumps.

The category names of categorical features were mapped to numbers, with 0 reserved for unknown values. Then, a classification tree was constructed for the 59400 samples, on a 60:40 random training testing data split, to predict the status of water pumps. The prediction accuracy on the test data was used as measure of performance of classification tree. In addition a new random forest was constructed including the installer of pump as an additional feature. Another random forest was constructed with 2 trees. The results of these three models were combined (ensembled), choosing the majority class predicted by the three models as the predicted output.

A random forest was generated to identify the relative importance of features used in the prediction. Random forest of varying sizes (from 2 to 25) were constructed and the accuracy for each of those were noted.

4. RESULTS

This preliminary analysis revealed that more than 70% of the variables are unknown (0 value) for the *amount of water available* variable. The univariate plots also reveals that *latitude* and *longitude* features contain a significant number of unknowns (coded as 0), and *year of construction* is not mentioned for a large number of pumps (coded as 0).

4.1 Descriptive Statistics

Tables 1 and 2 shows the descriptive statistics for the quantitative and categorical variables respectively. In the sample, *status of water pump* has an almost equal number of functional and non-functional pumps, and a relatively small number of pumps that require repair.

Table 1: Quantitative variables summary

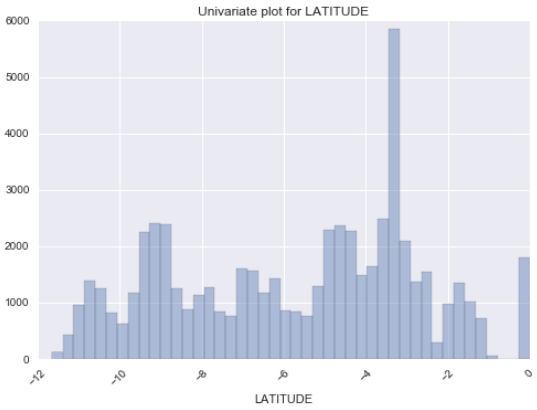
	25%	50%	75%	max	mean	min	std
AMOUNT OF WATER AVAILABLE	0	0	20	350000	317.65	0	2997.57
ALTITUDE OF THE WELL	0	369	1319.25	2770	668.801	0	692.621
POPULATION AROUND THE WELL	0	25	215	30500	179.91	0	471.482
LONGITUDE	33.0903	34.9087	37.1784	40.3452	34.0774	0	6.56743
LATITUDE	-8.54062	-5.0216	-3.32616	-2e-08	-5.70603	-11.6494	2.94602

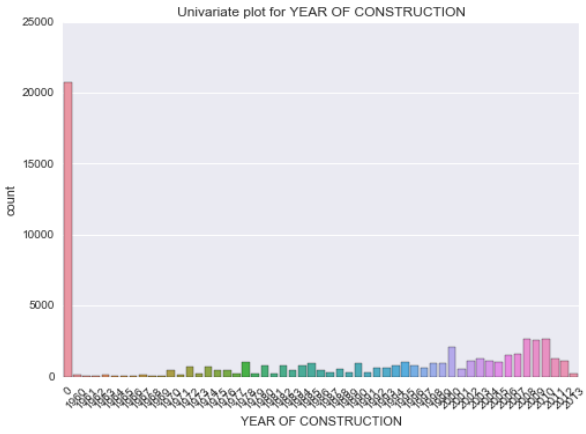
Table 2: Categorical variables summary

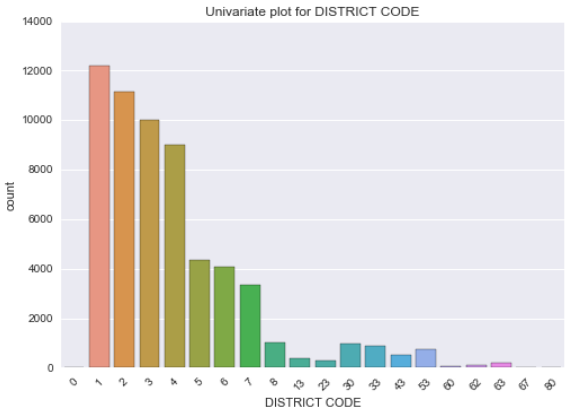
	count	freq	top	unique
YEAR OF CONSTRUCTION	59400	20709	0.0	55
GEOGRAPHIC WATER BASIN	59400	10248	lake victoria	9
REGION CODE	59400	5300	11	27
DISTRICT CODE	59400	12203	1	20
LGA	59400	2503	njombe	125
SCHEME MANAGEMENT	59400	36793	vwc	13
WATER EXTRACTION TYPE	59400	26780	gravity	18
MANAGEMENT	59400	40507	vwc	12
MANAGEMENT GROUP	59400	52490	user-group	5
PAYMENT FOR PUMP WATER	59400	25348	never pay	7
WATER QUALITY	59400	50818	soft	8
QUANTITY OF WATER AVAILABLE	59400	33186	enough	5
WATER SOURCE	59400	17021	spring	10
WATERPOINT TYPE	59400	28522	communal standpipe	7
PUMP INSTALLER	55745	17402	DWE	2145
WATER PUMP STATUS	59400	32259	functional	3

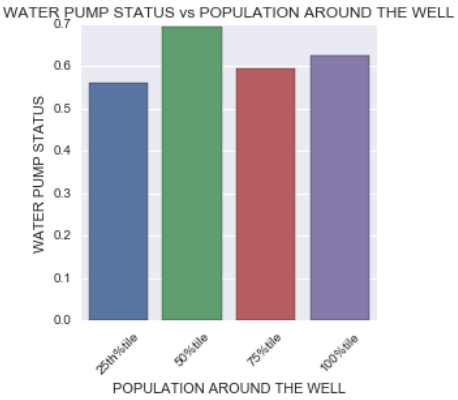
4.2 **Univariate Plots**

The distribution of values is easier to understand through visualizations. So, univariate plots were constructed. The plots of the quantitative variables are dominated by 0 values due to a large number of unknowns. Few of the plots are shown below.









4.3 **Bi-variate Plots**

The plots reveals that the percent of functional pumps increase with increase in amount of water. This association is expected.

As the depth of well increases, the percent of functional pumps steadily increase, implying a positive association.

Functioning of pump is more or less unaffected by the population surrounding the water pump.

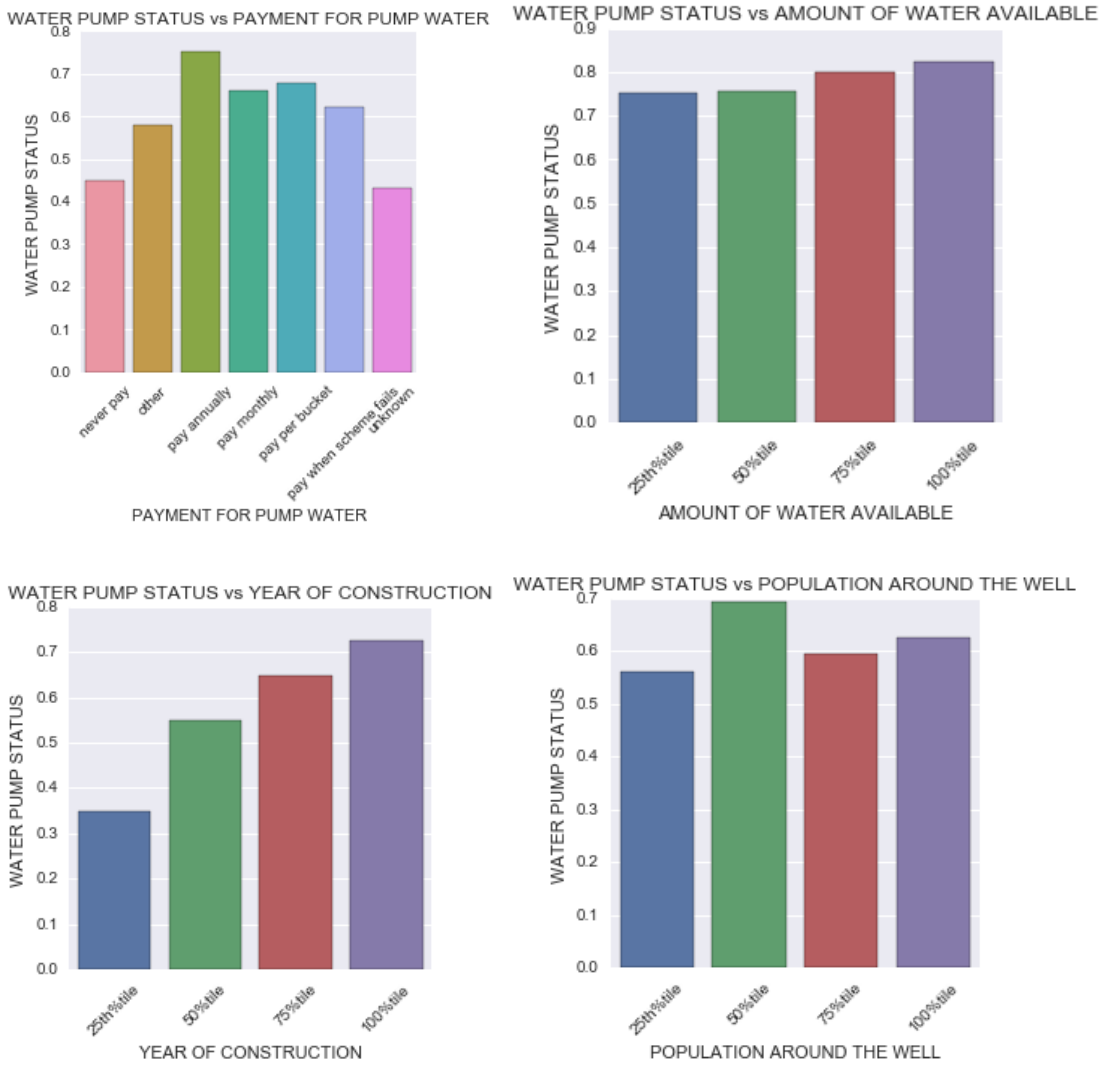
The plots for latitude and longitude reveal that some regions have comparatively better pumps functioning than others.

Functioning of pump is positively associated with payment. Pumps for which payment is collected (per bucket, per month or per year) show higher percent of functioning than for those which have no payment.

Pump for which people report as having enough quantity of water are functioning, compared to those which have insufficient water or are reported dry, signifying a positive relation. This association is expected.

Most recently constructed pumps are relatively more functional than those that are non-functional.

The rest of the variables have discrete categories which do not have a straight forward relationship with the response. Few of the bi-variate plots are shown below.



All the plots and code can be found at:

https://github.com/kishorepv/predict-water-pump-condition

4.4 **Chi-square Test**

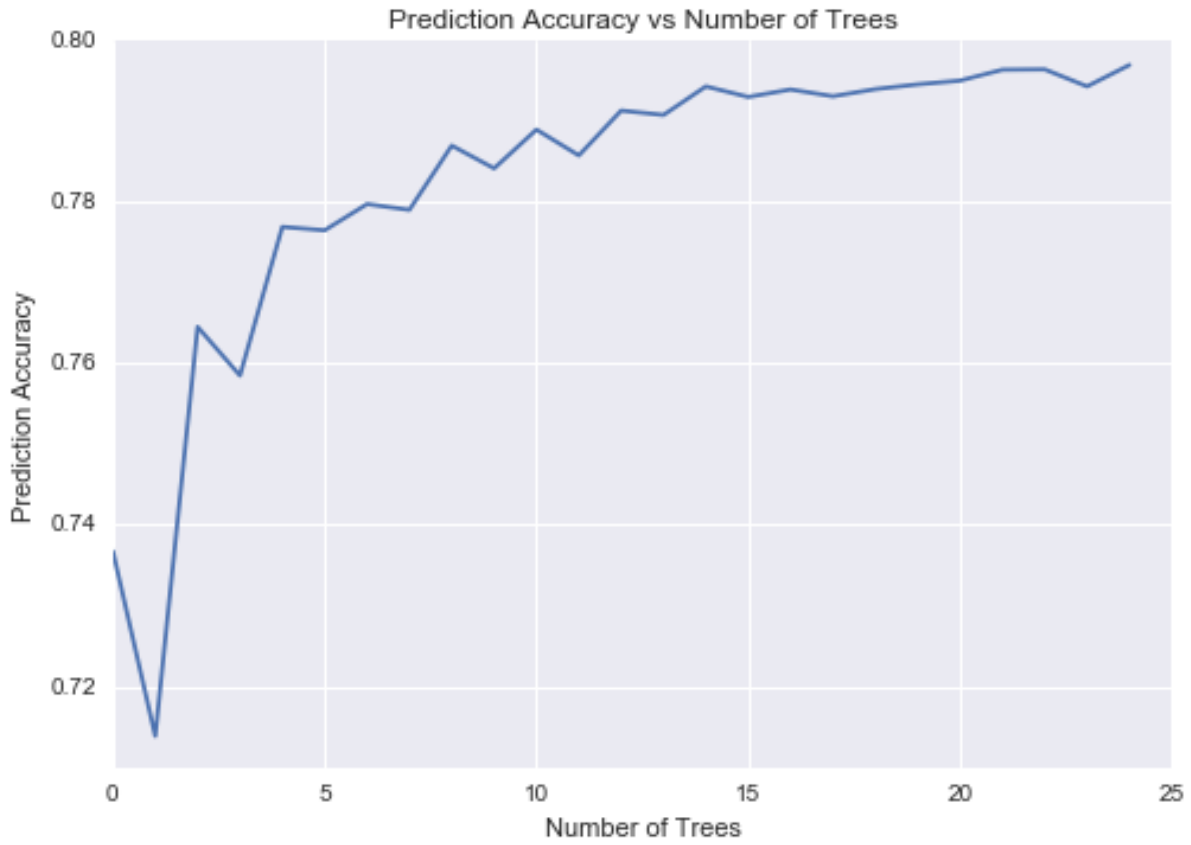
Table 3 shows the Chi-square test results. All of the features considered have a significant relation with the functioning of water pump (response variable) as evident by their p-values<0.05.
Further Post-Hoc test reveals that only some of the categories of the variables have a significant association with the response variable.

Table 3: Chi Square Test

	chi-square value	p value
YEAR OF CONSTRUCTION	3719.41	0
GEOGRAPHIC WATER BASIN	1579.77	0
REGION CODE	3303.8	0
DISTRICT CODE	1188.65	2.49625e-240
LGA	6155.39	0
SCHEME MANAGEMENT	1083.19	2.41198e-224
WATER EXTRACTION TYPE	5193.91	0
MANAGEMENT	1525.13	0
MANAGEMENT GROUP	135.412	2.70848e-28
PAYMENT FOR PUMP WATER	3169	0
WATER QUALITY	1518.6	0
QUANTITY OF WATER AVAILABLE	8625.38	0
WATER SOURCE	1185.14	1.95555e-249
WATERPOINT TYPE	6336.83	0
PUMP INSTALLER	1530.11	8.9123e-310
AMOUNT OF WATER AVAILABLE	130.529	4.15949e-28
ALTITUDE OF THE WELL	352.272	4.80424e-76
POPULATION AROUND THE WELL	208.478	6.21117e-45
LONGITUDE	413.188	3.07874e-89
LATITUDE	234.368	1.57191e-50

4.5 **Classification Tree and Random Forest**

The classification tree was trained on a random 60% of the dataset (N=59400) and tested on the remaining 40% of the data. The resultant normalized confusion matrix is shown in table 4. Table 6 shows the relative importance scores of features, produced by a random forest constructed for that purpose. The *latitude*, *longitude*, *quantity of water* and *type of waterpoint* are more important (in the same order) than rest of the other 15 features. Table 5 shows the normalized confusion matrix for the Random Forest classifier.



The graph of number of trees in random forest v/s prediction accuracy shows (initially jagged) increase in prediction accuracy (on test data), as the number of trees in the forest increase. It is shown in the pot above. Random forest with 20-25 trees provide the best prediction accuracies (0.795, 0.796, 0.797, 0.796, 0.796 and 0.798 respectively). A random forest with 2 trees provided a prediction accuracy of 11%. The negated prediction of this model was used with two other models- the discussed model above and a classification tree with inclusion of installer of pump as a feature. The ensemble model predicted the majority class of the predictions of these three models.

Table 4: Normalized Confusion Matrix (Classification Tree)

	non functional	functional needs repair	functional
non functional	0.76	0.19	0.15
functional needs repair	0.04	0.37	0.06
functional	0.2	0.44	0.79

Table 5: Normalized Confusion Matrix (Random Forest)

	non functional	functional needs repair	functional
non functional	0.76	0.19	0.15
functional needs repair	0.04	0.37	0.06
functional	0.2	0.44	0.79

Table 6: Relative Importance of features

	Relative Importance
LONGITUDE	0.15
LATITUDE	0.14
QUANTITY OF WATER AVAILABLE	0.14
WATERPOINT TYPE	0.08
ALTITUDE OF THE WELL	0.07
YEAR OF CONSTRUCTION	0.05
POPULATION AROUND THE WELL	0.05
PUMP INSTALLER	0.04
PAYMENT FOR PUMP WATER	0.04
WATER EXTRACTION TYPE	0.04
WATER SOURCE	0.03
DISTRICT CODE	0.03
REGION CODE	0.03
GEOGRAPHIC WATER BASIN	0.02
SCHEME MANAGEMENT	0.02
WATER QUALITY	0.02
MANAGEMENT	0.02
AMOUNT OF WATER AVAILABLE	0.02
MANAGEMENT GROUP	0.01

5. CONCLUSIONS/LIMITATIONS

In this project a classification tree was constructed to predict the functionality of a water pump in Tanzanian communities, given a set of features. A subset of the provided features in the dataset were utilized for the prediction. Using an ensemble of different models the prediction accuracy was improved. The goal of this study being predicting the functioning of water pumps, has been achieved.

The random forest constructed shows that not all of the features are strong indicators of functioning of the water pump. The *latitude*, *longitude*, and *quantity of water available* features are more relevant than others. These features alone can predict the working condition of water pump with an accuracy of 69.3%. With all the features taken together, the accuracy improves to 74.6%. Using an ensemble of three models, prediction accuracy was further improved to 80.01%.

These predictions can aid the water pump management be vigilant in their tasks of managing and improving the availability of drinking water to Tanzanian communities. Predicting the state of a water pump can help the maintenance teams to carry out their repair operations in an effective manner. They can prioritize repairs of non-functional pumps over all other tasks, to ensure faster re-availability of water to the affected people. Repair of functional but defective pumps can be prioritized over regular maintenance checks, thus help prevent them from completely failing. High availability of water can in turn motivate people to provide accurate data about the water pumps in their community. Accurate data can help better predict pump failure.

One of the main problems faced while predicting was the unavailability of sufficient data for some important features like *amount of water available, latitude, longitude, depth and installer of pumps*. Many of these features had 0 for their value and some variables even had invalid values (for instance depth had many negative values). The *installer* variable requires user to enter the name of pump installer. A lot of names are misspelled, which if overlooked can result in treating each variation of name of installer as a separate installer. This can lead to erroneous prediction.

Some of the data was collected through forms voluntary filled by people. This brings to question the authenticity of the data provided. Inaccuracies might have crept in due to lack of knowledge of the volunteer or malicious intent of some people. For example- a pump installer can enter fake entries which specify self as the installer and mark all the pumps as fully functional. An algorithm that considers the installer as a feature, can give skewed results that can lead one to consider this installer as reliable. Or he/she could do the similar thing for competitors mentioning all those pumps as under repair or non-functional.

Handling unknown values is another problem. Many approaches are suggested for dealing with them. The accuracy of prediction is affected by how this problem is approached.

Also one should be aware that this is not an experimental study. As a result, all associations that are mentioned between the features and response variable (functioning of pump) does not imply a causal relationship. This algorithm cannot be used as it is to predict the state of water pumps in other communities, like a different county etc. The association between features and response variable discussed in this project can be pseudo-associations i.e. there can be unknown variables (confounding variables) that can explain the apparent relation. All possible confounding effects are not analysed in this project.

This project predicts the functioning of the pumps, given certain features. One possible extension of this project can be to predict the time when a given water pump would fail. This can help the management of water pumps to embrace a preventive approach to managing the pumps. Other prediction algorithms not considered in this project can be combined to create different ensemble of models, which can provide better prediction accuracy. Future work can also explore on using a different way of handling unknown data.