Prediction of Tanzanian water pumps functional status

Frederic Collonval

September 15th, 2016

This study will focus on identifying the functional status (functional, needs repair or non functional) of Tanzanian water pumps. The possible explanatory variables will be location, construction year, funder, type of extraction, water quality and quantity, population using it, management organization and payment methods.

I picked up this challenge from the DrivenData competitions list because it shows a direct and practical application of how statistical analysis can help improve services and products quality. And as an engineer, those goals will be definitely the basis of any data science case I will have to solve. Moreover, as lots of possible explanatory variables are available, this will give me the chance to apply advance tools I learned during the Data Analysis and Interpretation online Specialization.

Predicting accurately the water pumps functional status will help planning maintenance earlier. That in turn will increase the availability of the water point and thus the quality of life for the people depending on those water supplies.

Methods

Sample

The database contains 74,250 records of water points information from the Tanzania Ministry of Water. The records were made between October 2002 and December 2013. Unfortunately there are no explanation on the techniques used to collect those data.

Measures

The functional status of the water points are categorized in three groups: functional, functional needs repair and non functional.

The potential predictors will be:

- The amount of water available; missing data are coded as 0, they will be replaced by the mean value to suppress minimum amount of data.

- The organization having funded the well

  From the various actors, the following categories will be created :

  ```
  ➔    'organisation' : ('bank', 'msf', 'wwf', 'unicef', 'unisef',
  'oxfam', 'oxfarm', 'rotary club', 'lion's club', 'care', 'without',
  'action contre la faim', 'rain', 'red cross', 'blue angels', 'fundat',
  'foundation'),
  ➔    'church' : ('church', 'churc', 'rcchurch', 'roman', 'missionsry',
  'lutheran', 'islamic', 'islam', 'vision'),
  ➔    'private' : ('consulting', 'engineer', 'private', 'ltd', 'co.ltd',
  'contractor', 'enterp', 'enterpr', 'company', 'contract'),
  ➔    'community' : ('village', 'community', 'communit', 'district',
  'council', 'commu', 'villigers', 'villagers'),
  ➔    'government' : ('government', 'gov', 'govt', 'gover', 'gove',
  'governme', 'ministry'),
  ➔    'other' : ('0', 'nan', 'known', 'other', 'unknown'),
  ➔    'danida' : ('danida', 'danid'),
  ➔    'foreign government' : ('netherlands', 'germany', 'european')
  ```

  Then the 9 most funders will be kept and the others will be gathered in the `other` category.

- The installer of the well; the grouping technique applied on the funders will be applied on the installer categories.
- The GPS coordinates (height, longitude and latitude); missing data are coded as 0, they will be coded as NaN except for the height for which the missing data will be replaced by the mean values to keep a maximum of records for the analysis.
- The geographic water basin
- The geographic region
- The population around the well; missing data are coded as 0, they will be coded as NaN.
- Organization of public meeting at the water point; dichotomous variable (True/False)
- The operator of the water point
- The management of the water point
- Does the water point receive a construction permit?
- Year the water point was constructed; missing data are coded as 0, they will be replaced by the median value to avoid discarding a lot of records in the analysis.
- The kind of extraction the water point uses
- How payment are handled?
- The quality of the water
- The quantity of the water
- The source of the water
- The type of water point

As the Python package `sklearn` cannot handle non-binary categorical variables, those variables will be expanded in as much new dichotomous variables as there are categories. Therefore the number of potential explanatory variables will be huge. So as a prepocess steps, a random forest test will be carried out to select only the variables having a substantial effect.

Analyzes

The distributions of the response and explanatory variables will be evaluated by looking at the frequency tables for categorical variables and by calculating statistical values (mean, standard deviation, minimum and maximum) for quantitative variables.

The response variable being categorical, bivariate associations will be visualized using bar charts after collapsing categories if needed. And the possible bivariate associations will be tested using Chi-Square test.

The random forest method will be applied to identify the best subset of predictors. The DrivenData competition has split the database in a training set containing 80% of the records and 20% are kept for testing by submission on the website. As multiple submissions are allowed for the competition, the accuracy of the model will be tested by submitting the prediction carried out on the test data.

Results

Explanatory variable selection

First a random tree test was performed to limit the number of explanatory variables. From that first analysis (see the Table 1 below), the following explanatory variables are kept:

- The gps coordinates – longitude, latitude and height – of the water point
- The quantity of water available
- The population size next to the water point
- The year of construction
- If a permit was issued or not for the water point
- The type of extraction
- The water point type
- The payment methods

Although GPS coordinates are important, the administration division (like geographic region) has low importance. It seems also than the way the water point was funded and installed and how it is managed are not of great importance. Some natural guesses like the quantity, the population living around and the year of construction come forward in the random forest test.

	importance
longitude	0.140741
latitude	0.137381
dry	0.083944
height	0.070537
population	0.047320
construction year	0.045558
enough	0.029084
other extraction type	0.027418
other water point type	0.023971
never pay	0.017196
insufficient	0.014567
water_amount	0.013900
permit	0.011996
gravity	0.010515
other funder	0.010112

Table 1: The 15 most important features in the dataset.

Descriptive Statistics

In the training data set, 54.3% (N=32259) of the water point are functional, 7.3% (N=4317) need repair and 38.4% (N=22824) are non functional.

For those water points, the quantity of water available is *enough* for 55.9% (N=41522), *insufficient* for 25.4% (N=18896) and *dry* for 10.5% (N=7782). The quantity is unknown for 1.3% of the data (N=975).

The majority of the point are communal standpipes (58.2%, N=43239). The second most important type is hand pump type (29.5%, N=21884).

The method to extract the data are mostly gravity (44.8%, N=33263) and hand pumps (27.7%, N=20612).

To get water, people are usually never paying (42.7%, N=31712). For the points for which people pay, they are doing so on bucket basis (15.2%, N=11266) or by recurrent payment; monthly for 14% (N=10397) or annually for 6.1% (N=4570). The payment method is unknown for 13.7% of the cases (N=10149).

The majority of the water points were constructed with a permit (65.4%, N=48606). But 29.4% (N=21851) were not built having one. And the permit status is unknown for 5.1% of the water points (N=3793).

The distribution of the quantitative variables are presented in the Table 2 below.

	count	mean	std	min	max
construction year	38691	1996.8	12.472	1960	2013
height	38962	1018.9	612.57	-90	2770
latitude	59400	-5.706	2.946	-11.649	-2e-08
longitude	57588	35.15	2.6074	29.607	40.345
population	38019	281.09	564.69	1	30500

Table 2: Quantitative variable distributions.

Bivariate analyzes¶

The figures below show the mean value of the *functional* variable (0 = non functional, 1 otherwise) for the different categorical variables.

Using post hoc chi-square tests, the major conclusions drawn are:

- Water points working with gravity have significantly more chance to be functional (max(p-value) = 1.4 < 0.05/21). And non-mentioned extraction are the more likely to be non functional.
- Water points type *cattle trough* and *improved spring* have no significant differences. And they are the two types having the highest probability to be functional. No conclusion can be drawn for the *dam* type as only 5 functional points are reported. The water points of type *other* are the most likely to be non functional.
- Water points for which people are paying annually are the most likely to be functional. And the one free of charges or of unknown payment method are not significantly different and both have 50% chances to be non functional.
- *Dry* water points are most likely to be non functional. And those with *enough* or *seasonal* water are not significantly different and are the more likely to be functional.
- Water points having a construction permit have a significantly more chance to be functional than those not having a permit (p-value = 1.4e-26).

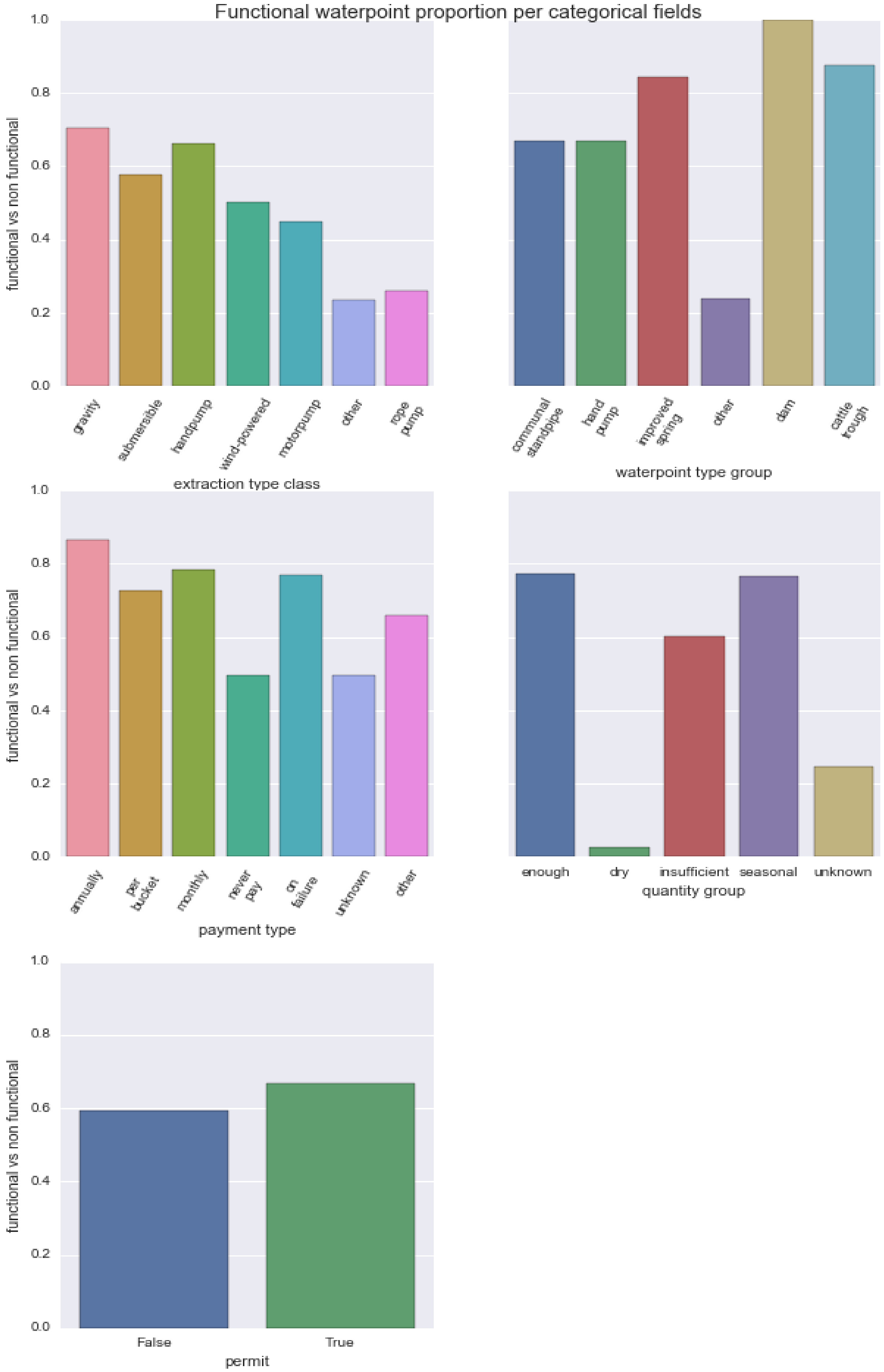

Figure 1: Relationship between the categorical variables and the functional status of the water points.

To visualize the influence of the quantitative variables on the functional status of the water points, the quantitative variables have been collapsed in two bins; the median value being the separation.

Using chi-square test, all variables have a significant relationship with the response variable. Water points with higher altitude are more likely to be functional (p-value = 2e-57). Those more in the eastern side of Tanzania have a lesser chance to be functional (p-value = 0.003). The water points constructed after 2000 are in better functional condition (p-value = 0). And those sustaining higher population tend to be less functional (p-value = 2.5e-13).

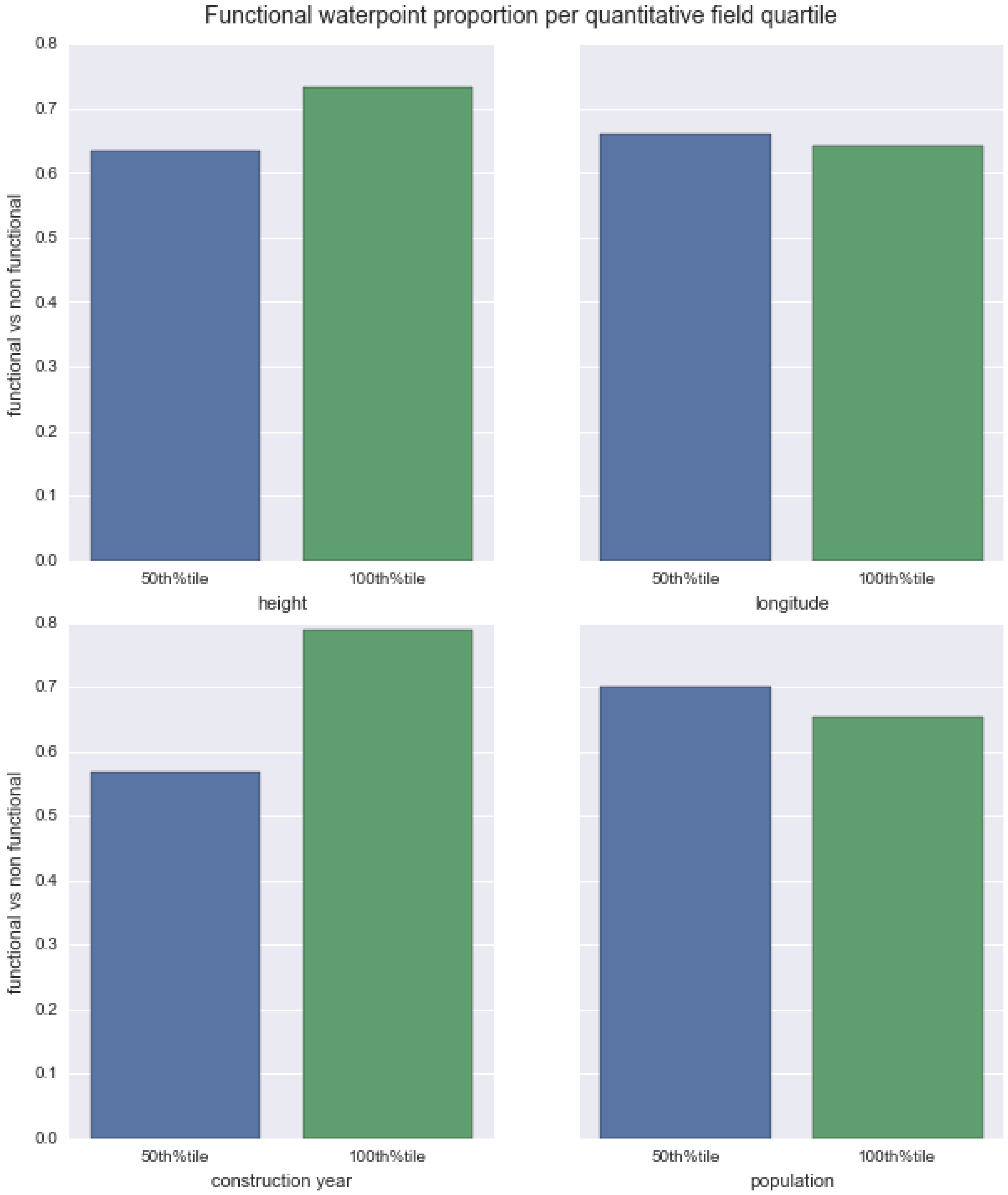

Figure 2: Bivariate relationship between the quantitative variables (binned in two categories around the median) and the functional status of the water points.

Random Forest Test

With the subset of explanatory variables selected, we can split the data to estimate the number of trees needed to stabilize the accuracy. By taking 60% of the available data as training set, the accuracy of the random forest test stabilizes for a number of trees superior to 23 as shown in the figure below.

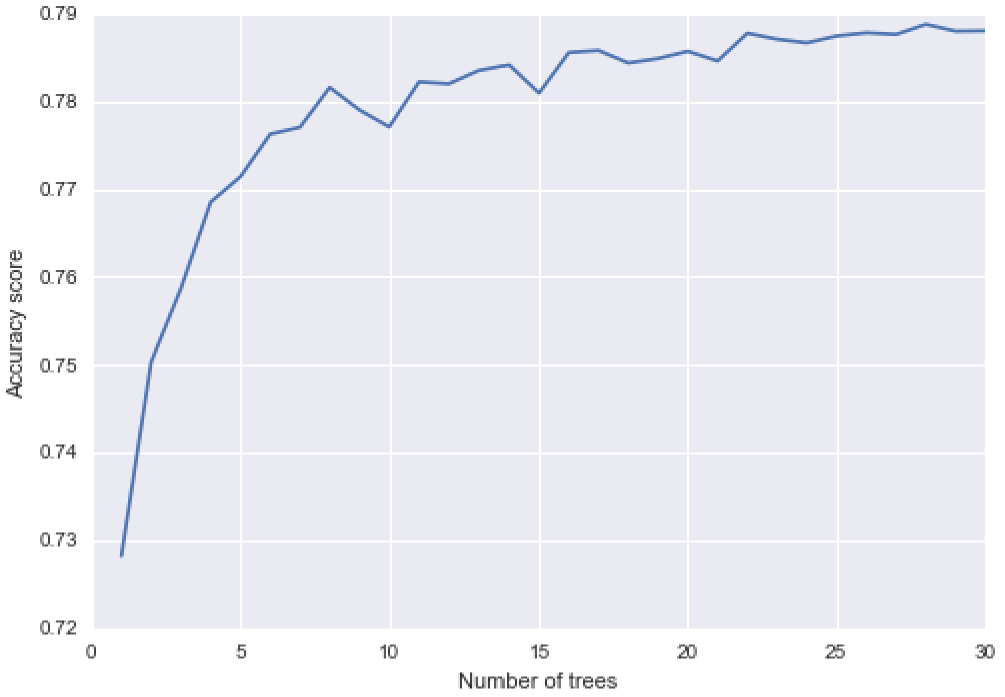

Figure 3: Accuracy score evolution with the number of trees used in the random forest test.

So I run a random forest test with 25 trees with all training data and submitted on DrivenData.org the resulting prediction. I got an accuracy score of 76.86%.

Conclusion

This project used random forest test to identify the variables influencing the most the functional status of Tanzanian water pumps from N=74250 water points characteristics recorded between October 2002 and December 2013 by the Tanzanian Ministry of Water. There are around 55% of pumps working properly, 7% in needs of repair and 38% non functional.

Applying the random forest test, the number of potential explanatory variables was reduced from 20 to 10 by looking at the importance of each features. The most influential variables are the GPS coordinates (longitude, latitude and height). Then comes the quantity of water available, the population living around the pumps, the type of extraction and the year of construction.

The random forest test using 25 trees had an accuracy score of 76.9% when tested against the DrivenData test set. The optimal number of trees was found by optimizing the accuracy score with the number of trees after dividing the provided data in two groups; 60% to train the method and 40% to test it. As the best score obtain was around 78.9%, it can be said that the model will predict fairly well new dataset.

From the feature importance calculation, it can be concluded that an improved water reparation policy should focus on dispatching teams not evenly in the country as the GPS coordinates influence greatly the water pumps status. And the primarily target should be based on the population size living around the water point and its year of construction.

Although lots of parameters have been recorded for this analysis, it is possible that a factor non considered here is important and is confounding other factors reported here.

From the analysis, the funder and the installer do not seem to have a big impact on the functional status. But as those two categories contain a wide variety of answers (some containing spelling mistakes or abbreviations), a deeper analysis of those two categories should be carried out to gather in meaningful categories the various actors. Right now some doubts remain on a potential confounder effect. Some parameters statistically important (population, height and construction year) have lots of missing data. In this study, the missing data of those variables were filled by their mean or their median values to avoid dropping to many records. Trying to fulfill the missing data will help improving the accuracy. Therefore, adding additional records and fulfilling the missing value should be the priority of any additional effort to improve the predictive algorithm.

> The Jupyter notebook used to generate this final report is available there: https://github.com/fcollonval/coursera_data_visualization/blob/master/WaterPumpsPrediction.ipynb.