Prediction of Tanzanian water pumps functional status

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 and management organization.

I picked up this challenge from the DrivenData competitions list because it shows a direct and pratical 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 learnt 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 clear 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. As the response variable contains more than 2 categories, two new dichotomic variables will be created by collapsing the three existing categories :

- *functional* : 1 for functional and functional needs repair waterpoints and 0 for non functional waterpoints.
- *no repair* : 1 for functional waterpoints and 0 for the others.

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; dichotomic variable (True/False)
- The operator of the waterpoint
- The management of the waterpoint
- Is the waterpoint is permitted?
- Year the waterpoint 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 waterpoint 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 dichotomic 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.

Analyses

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 reponse 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 below), the following explanatory variables are kept:

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

The primary consequences is althought gps positions are important the administration division (like geographic region) have less importance. It seems also than the way the waterpoint was funded and installed and how it is managed are not of great importances. Some natural guesses like the quantity, the population living around and the year of construction come forward in the random forest test.

	importance
longitude	0.139462
latitude	0.139410
dry	0.093651
height	0.069156
population	0.047748
construction_year	0.046577
other_waterpoint_type_group	0.039365
other_extraction_type_class	0.023615
enough	0.022052
never pay	0.020961
water_amount	0.013770
permit	0.013237
government_funder	0.012220
insufficient	0.011966
dwe	0.010880

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

Descriptive Statistics

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

For those waterpoints, 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 waterpoints 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 waterpoints (N=3793).

The distribution of the quantitative variables are presented in the table below.

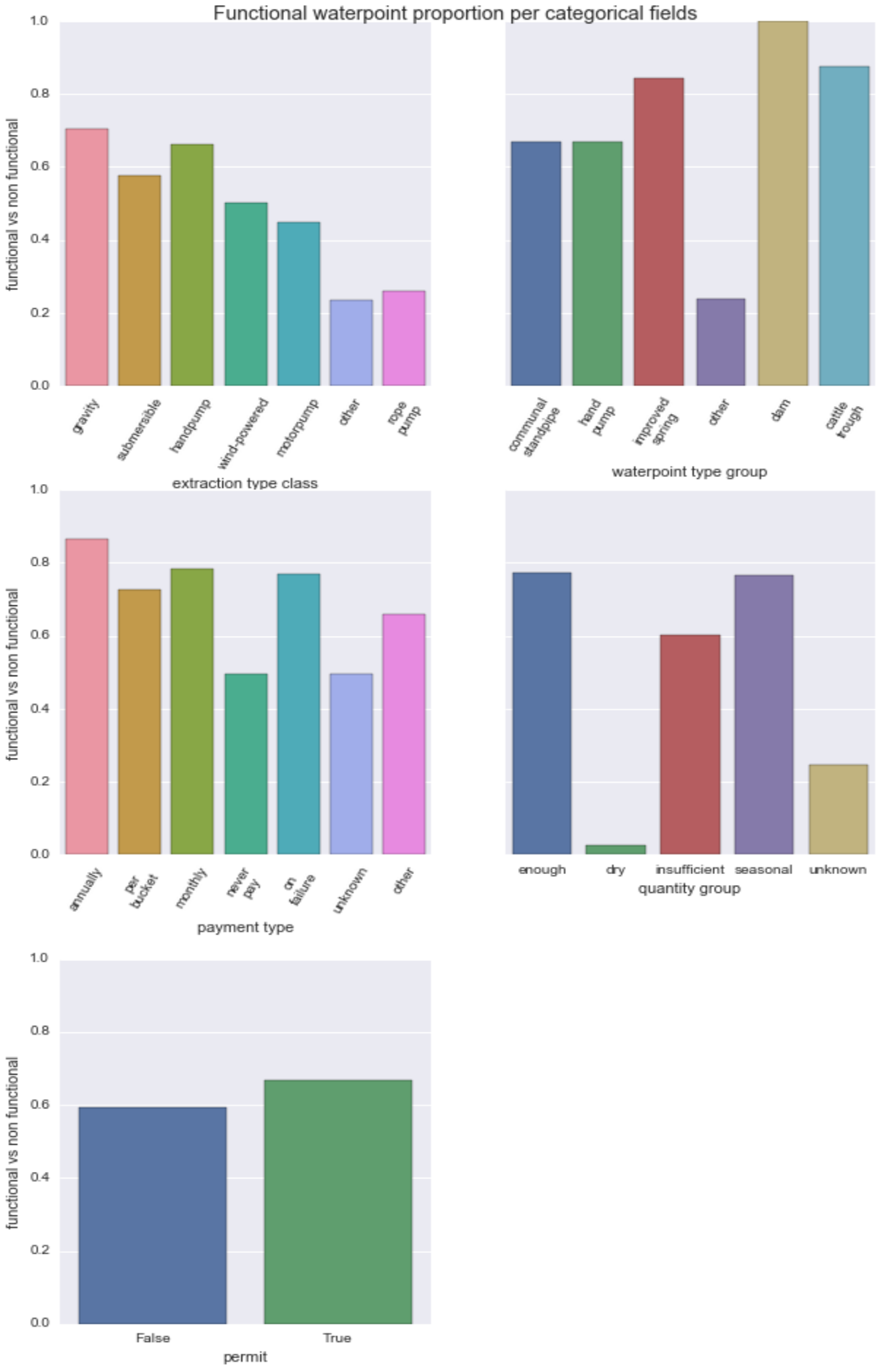
	count	mean	std	min	max
construction year	38691	1996.8	12.472	1960	2013
height	38962	1018.9	612.57	-90	2770
latitude	59400	-5.706	2.946	-11.649	-2e-08
longitude	57588	35.15	2.6074	29.607	40.345
population	38019	281.09	564.69	1	30500

Bivariate analyses

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

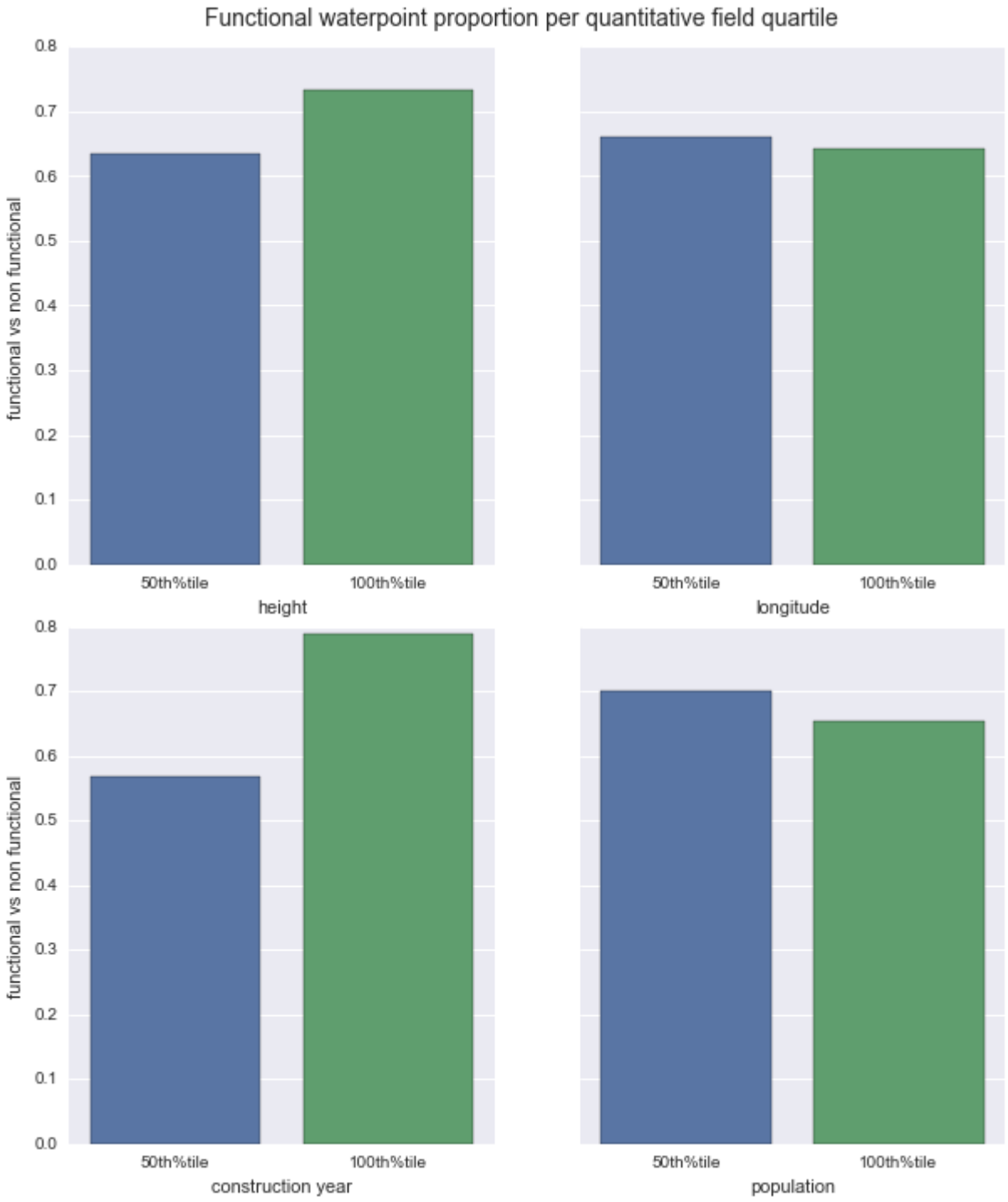
Using post hoc chi-square tests, the major conclusions drawn are :

- Waterpoints 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.
- Waterpoints 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 waterpoints of type *other* are the most likely to be non functional.
- Waterpoints for which people are paying annualy 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* waterpoints 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.
- Waterpoints having a construction permit have a significantly more chance to be functional than those not having a permit (p-value = 1.4e-26).


Functional waterpoint proportion per categorical fields
functional vs non functional
1.0
0.8
0.6
0.4
0.2
0.0
gravity
submersible
handpump
wind-powered
motorpump
other
rope pump
extraction type class
communal standpipe
hand pump
improved spring
other
dam
cattle trough
waterpoint type group
annually
per bucket
monthly
never pay
on failure
unknown
other
payment type
enough
dry
insufficient
seasonal
unknown
quantity group
False
True
permit

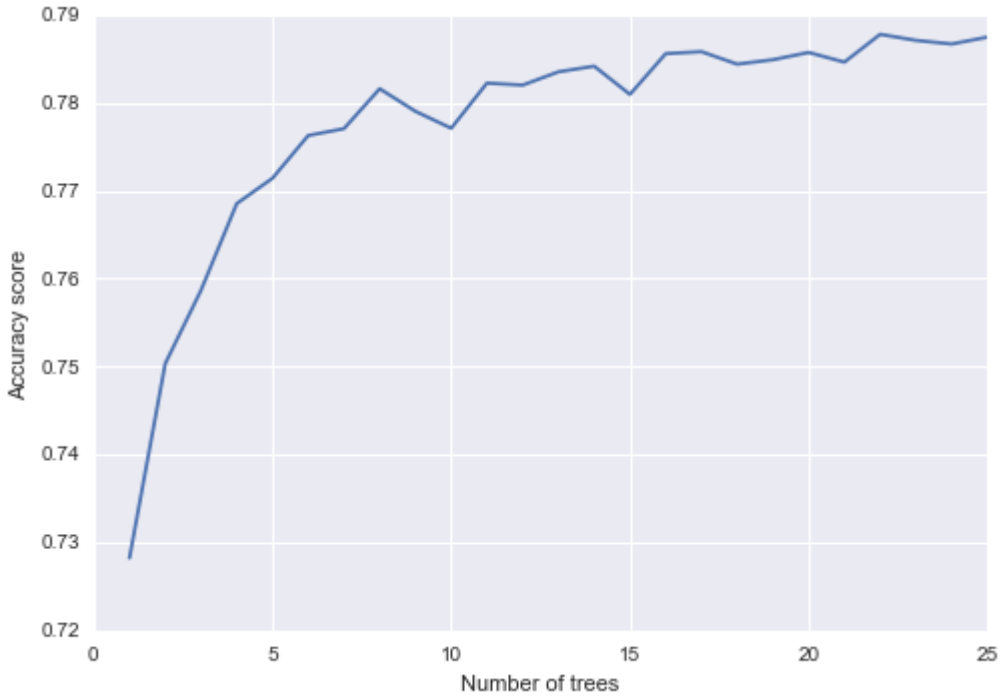
To visualize the influence of the quantitative variables on the functional status of the waterpoints, the quantitative variables have been collapsed in two bins the median value being the separation.

Using chi-square test, all variables have a significant relationship with the response variable. Waterpoints with an 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 waterpoints constructed after 2000 are in better functional condition (p-value = 0). And those sustaining an higher population tend to be less functional (p-value = 2.5e-13).



Random Forest Test

With the subset of explanatory variables selected, we can split the data to estimate the number of trees needed to stabilise 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.



So I run a Random Forest test with 25 trees with all training data and submitted on DrivenData.org the resulting prediction. I got a score of 76.86%. That score is a bit frustrating as the majority of the people reach 80%. So I'll work hard during the last week to improve the results.

> This Jupyter notebook will be the basis for the final report for the Data Analysis and Interpretation Specialization