**Pump it Up: Data Mining the Water Table**

Antriksh Agarwal

Sarvesh Pandit

**Abstract**

The task of this project is to predict which water pumps throughout Tanzania are functional, which of those need repairs, and which do not work at all. The prediction is based on a number of variables about what kind of pump is operating, when it was installed, and how is it managed. We are comparing the performance of a couple of boosted classifiers with Support Vector Machines (SVMs) and the performance of SVMs with different kernel functions. The selected boosted classifiers for this project will be XGBoost and LightGBM. The ultimate goal would be to find the best classifier and suggest improvements.

# Introduction

The project aims to identify problems with various factors that affect the functioning of water pumps, how to use these factors to an advantageous outcome and how to overcome the problems of these factors. This is to be done by gaining a smart understanding of the features because of which the water pumps installed by various organizations in Tanzania have stopped functioning properly.

# Related Work

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts. True Type 1 or Open Type fonts are required. Please embed all fonts, in particular symbol fonts, as well, for math, etc.

# Dataset

The dataset for the task has been provided in the challenge Pump It Up: Data Mining the Water Table [1] posted on the Driven Data [2] website.

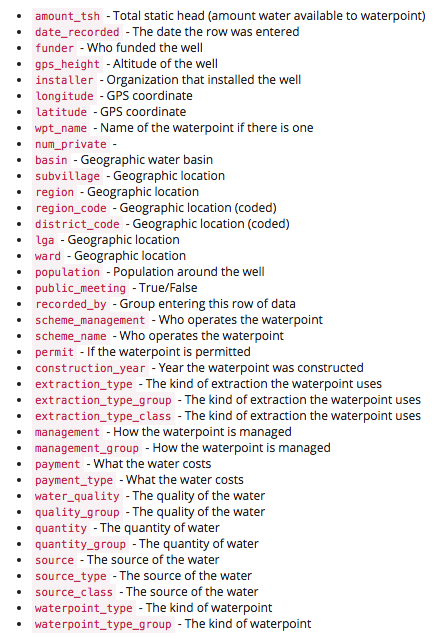
The features provided in the dataset have been described on the challenge webpage and are shown with their description in Fig. 2. There are 40 columns in the training dataset with over 59,000 instances of data. The labels for each of the instances have been provided in a separate file. Each of these instances have a classifying label which classifies the data point to either “functional”, “non functional” and “functional needs repair”. The labels mean the following:

* functional - The waterpoint is operational and there are no repairs needed
* functional needs repair - The waterpoint is operational, but needs repairs
* non functional - The waterpoint is not operational



Pump It Up: Data Mining the Water Table

We have 40 features in our dataset, the names and their description are given below.



Feature names and their descriptions as provided on the challenge website

# Experimental Methodology

For creating training, validation and test dataset, our plan is keep 20% of the data as the test data and for the remaining 80% of the data we will use 10 fold cross validation technique for generating training and validation datasets.

# Coding language / Technique to be used

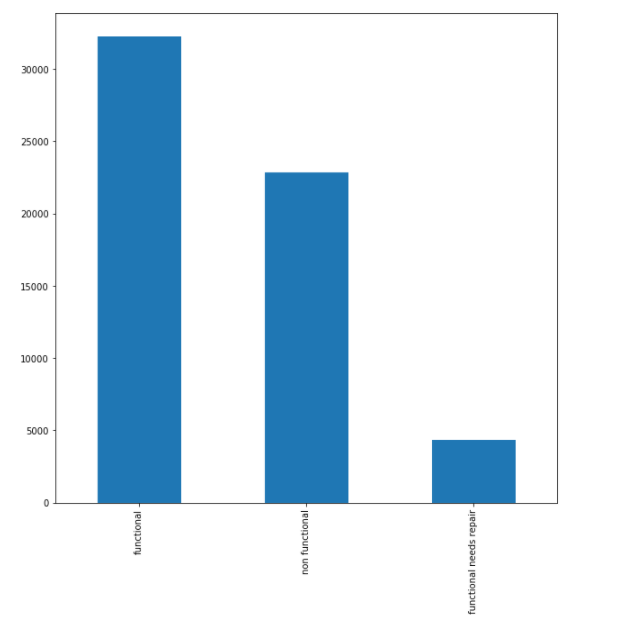
Coding language : Python

Techniques: SVM, XGBoost, LightGBM

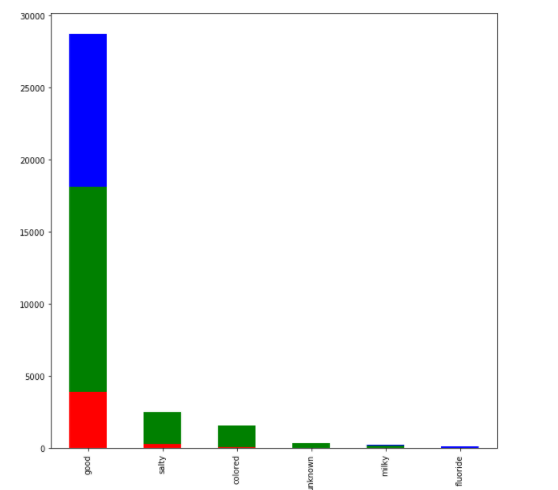
# Preliminary Result

# The dataset provided have 40 columns and 59,400 rows/instances. To have a brief understanding of the dataset, we decided to plot the data on a bar graph. We obtained the following results.

1. Out of the three labels (functional, non-functional, functional needs repair), over 30,000 instances have label as “functional”, about 23,000 instances have label as “non-functional” and a little less than 5,000 instances have label as “functional needs repair”



1. For “quality\_group” which is one of the feature, looking at the result, we came to a conclusion that if the value for quality\_group attribute is “good” the chances of it to be classified with the label “functional” are more whereas the chances of it to be classified with the label “non- functional” are more for rest of the values.

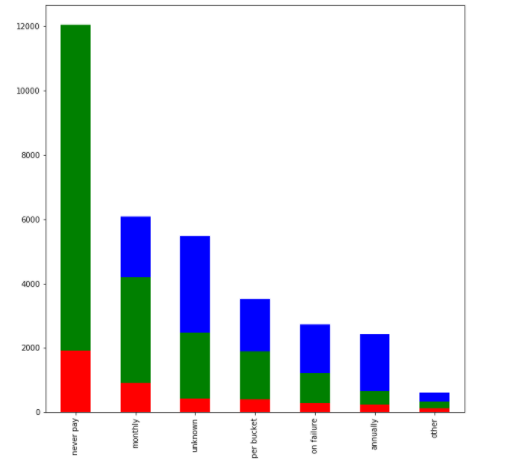


Blue color: functional

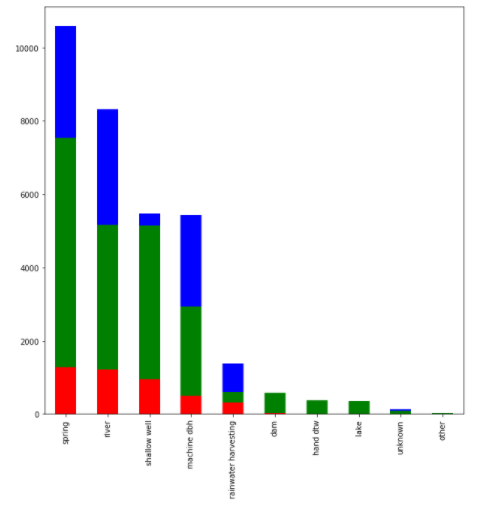
Green color: non-functional

Red color: functional needs repair

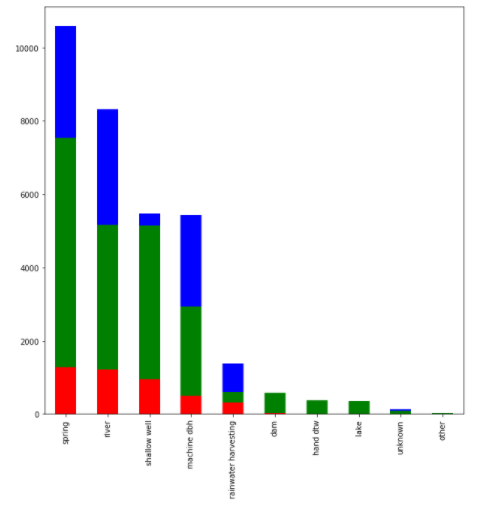
1. For the feature “payment\_type”, if the value for the this attribute is “never pay” then there is a good chance that the instance will be classified as “non-functional”, for all the other values the instance to be classified as “functional” has a greater chance.



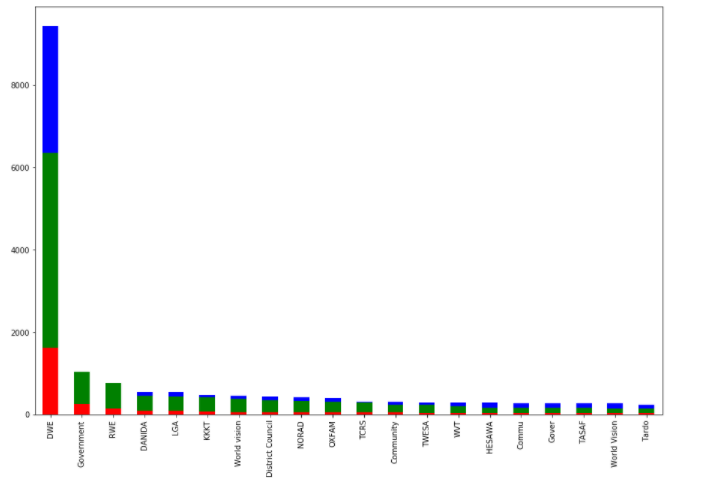
1. For the attribute “source”, the bar graph obtained was much more complex than the rest of the attributes and therefore, to derive a conclusion only by looking at the bar graph was not possible



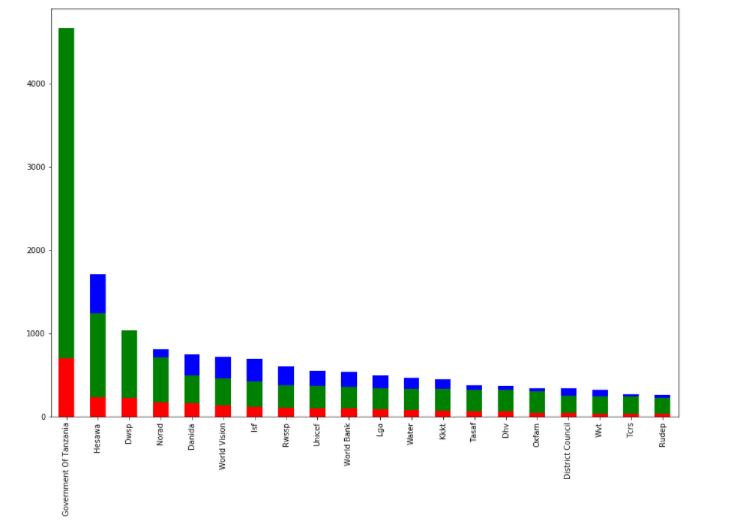
1. For the feature “quantity”, if the value of the feature is “enough” then it will have higher chances of it to be classified as “functional”, else the chances of it to be classified as ”non-functional” will be more.



1. For installer attribute, There are categories with similar names, but which are recognized different due to different cases (lowercase/uppercase, e.g. World vision and World Vision) of some part or whole of the category or some abbreviations (e.g. Gover and Government, Commu and Community). Our assumption is that if the name (in it's exact given form) is different, then the categories are different, else they are already being recognized as same. These are the results which we obtained for installer attribute.



1. For all the attributes related to geographic location (basin, subvillage, region, lga, ward), we combined them and obtained the following graph



# References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example: [1]. Where appropriate, include the name(s) of editors of referenced books. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in “[3]”—do not use “Ref. [3]” or “reference [3]”. Do not use reference citations as nouns of a sentence (e.g., not: “as the writer explains in [1]”).

Unless there are six authors or more give all authors’ names and do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/

https://www.drivendata.org/

J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

K. Elissa, “Title of paper if known,” unpublished.

R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.