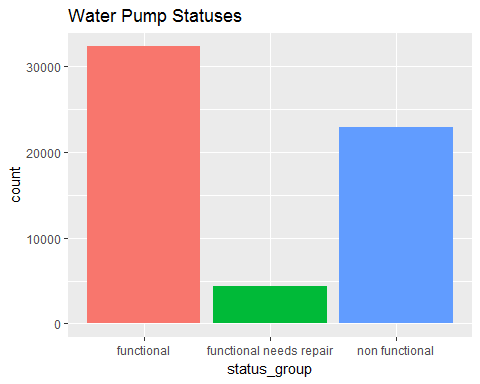
DrivenData: Pump it Up: Data Mining the Water Table

This is a project from DrivenData.

From the host: *Using data from Taarifa and the Tanzanian Ministry of Water, can you predict which pumps are functional, which need some repairs, and which don't work at all? A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.*

Water pumps can be in one of three states:

|  |  |  |
| --- | --- | --- |
| status\_group | count | percent |
| functional | 32259 | 0.5430808 |
| functional needs repair | 4317 | 0.0726768 |
| non functional | 22824 | 0.3842424 |



Code can be viewed in model.R file. Just presenting a few results here. A first attempt with logistic regression gives the following classification rate

predicted vs. actual

|  |  |  |  |
| --- | --- | --- | --- |
|  | functional | functional needs repair | non functional |
| functional | 14382 | 1607 | 3812 |
| functional needs repair | 74 | 186 | 80 |
| non functional | 1642 | 394 | 7523 |

## [1] 0.7438047

A second model, using a random forest, gives this out-of-sample classification rate

predicted vs. actual

|  |  |  |  |
| --- | --- | --- | --- |
|  | functional | functional needs repair | non functional |
| functional | 14575 | 1418 | 3127 |
| functional needs repair | 160 | 429 | 117 |
| non functional | 1363 | 340 | 8171 |

## [1] 0.780303

Here's how these two models would stack up to other submissions. Not a bad start for a few hours work, but looks like I've got some progress to make!

