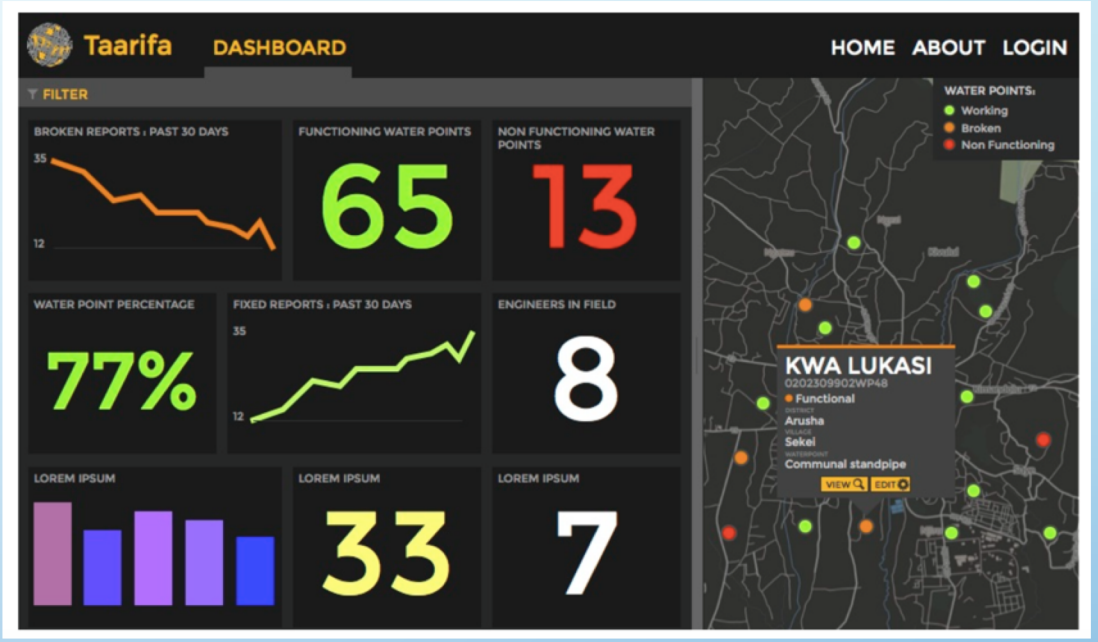
Using Machine Learning To Predict Water Well Function In Tanzania

Gina Durante
Flatiron School
Data Science Boot Camp
July 30, 2020

The problem: Predicting water wells in need of repair or replacement

- Almost 75,000 wells in the country; just over 54% are functional
- Water supply dashboard project with non-profit ***Taarifa*** → database of all water wells in the country
- Machine learning models may help us predict well status (DrivenData hosting machine learning competition)



What kinds of data are we talking about?

- Location
- Project technical specifics
- Project management specifics

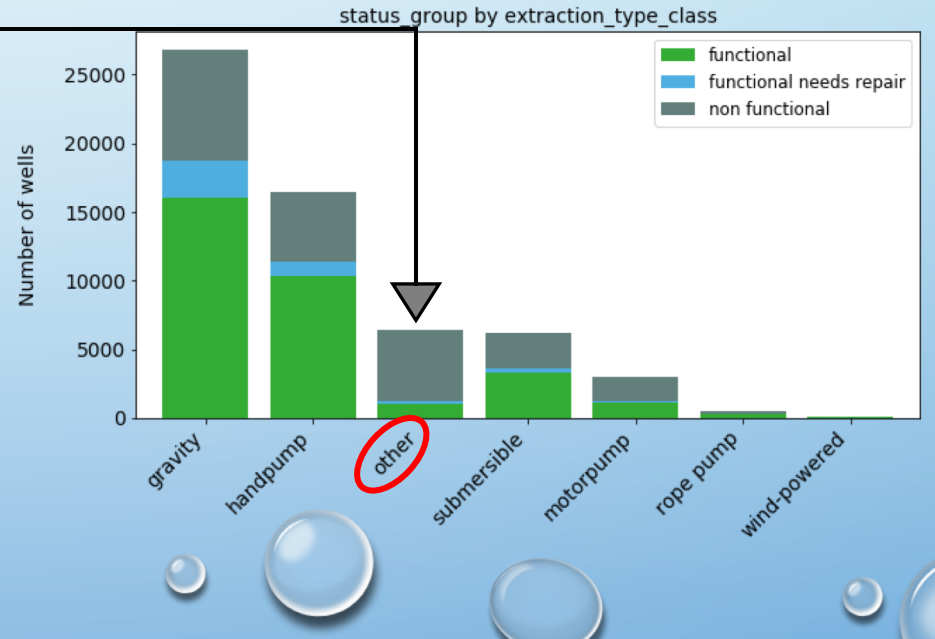
A Few Words on Methodology…

- Key elements of project:
  - Obtaining and 'cleaning' data
  - Data exploration and visualization
  - Running machine learning models, modifying parameters, and running models again
- Want a model that will:
  - Accurately predicts each well type ('functional', 'non-functional', or 'functional-needs repair')
  - Minimizes costly errors

Findings: Well type abundance and functional status

- The good news: three most abundant well types have highest functional rates
  - Gravity
  - Handpump
  - Submersible
- The not-so-good news:
  - ~84% (5,400) of extraction type 'other' wells aren't working
  - Only 38% of motorpumps are functional (but comprise a small fraction of wells--5%)
  - Rope pumps:
    - Highest functional percentage of any type (65%), but…
    - Comprise only 0.8% of all wells

Well Type	Functional percentage of *this type* of well	Percentage of *all* wells made up of this type
gravity	60%	45.1%
handpump	63%	27.7%
motorpump	**38%**	5.0%
other	**16%**	**10.8%**
rope pump	65%	0.8%
submersible	54%	10.4%
wind-powered	43%	0.2%
All wells	54.30%	100.0%



Findings: Variables Influencing Prediction Results

- Top 3 account for ~52%
  - LGA
  - Funder
  - Water abundance (quantity)
- Next 6 features account for ~45%
  - *Waterpoint type* (e.g., hand pump, communal standpipe, dam)
  - *Source* (e.g., river, spring, shallow well, rainwater harvesting)
  - *Extraction type* (e.g., gravity, hand pump, motor, submersible)
  - *Payment type* (e.g., by bucket, monthly, never)
  - *Management* (type of organization)
  - *Water quality* (e.g., soft, cloudy)

Variable	Ranking	Importance %*
LGA	1	25.0%
Funder	2	14.6%
Water abundance	3	12.0%
Waterpoint type	4	9.7%
Source	5	9.1%
Extraction type	6	8.7%
Payment type	7	8.2%
Management	8	5.6%
Water quality	9	3.8%
Permit	10	1.8%
Public meeting	11	1.5%

* Random Forest model, class_weight=balanced

Recommendations

- Evaluate the conditions for those specific variable values (e.g., LGAs, funder/installer, water source) highest on list of importances (see list in Appendix slides)
  - LGAs: Bariadi, Kigoma Rural, Kongwa, Makete, Siha, Rombo, Njombe, Chunya, Kyela, Rungwe
- Well extraction types
  - Discourage well extraction type **"other"** (only 16% are functional; 81% non-functional!) *unless* evidence for likely successful operation is provided
  - Investigate possible improvements for **motorpump** performance (esp. "cemo") (38% are functional; 57% are non-functional)
  - Explore increasing installation of **rope pumps** (65% functional; <1% of all wells)
- Water source
  - Investigate why lake water source have poor performance (only ~21% functional)
  - Evaluate whether dam water source performance can be improved (~38% functional)

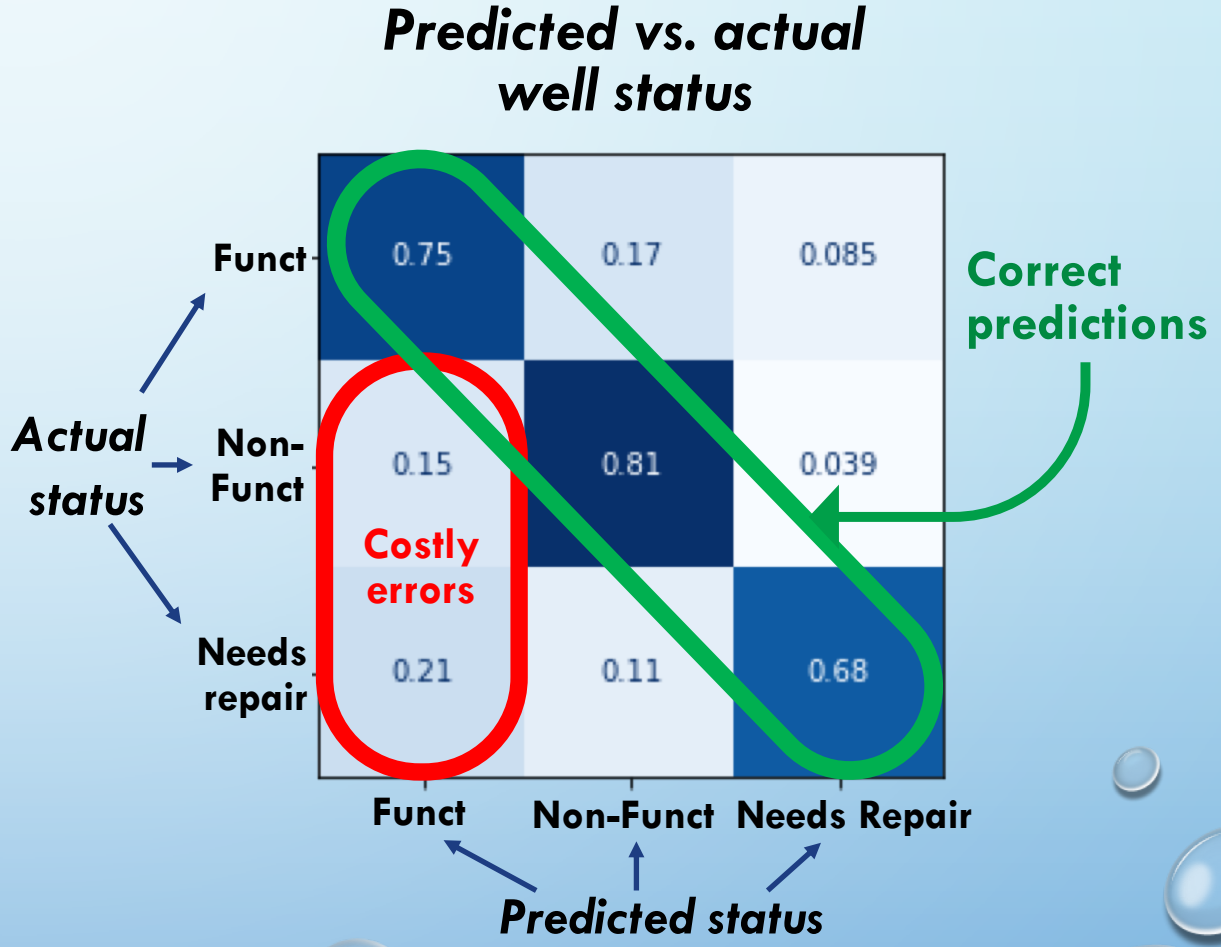
Question and Answer / Concluding Comments

- Q&A
- Thank you for your attention!

APPENDIX

A Few Words on Methodology…

- Key elements of project:
  - Obtaining and 'cleaning' data
  - Data exploration and visualization
  - Running machine learning models, modifying parameters, and running models again
- Want a model that will:
  - Accurately predicts each well type ('functional', 'non-functional', or 'functional-needs repair')
  - Minimizes costly errors



Features Consistently in Top 20 Importance Lists

- **LGAs:** Bariadi, Kigoma Rural, Kongwa, Makete, Siha, Rombo, Njombe, Chunya, Kyela, Rungwe
- **Funders:** govt_of_tanzania, dwsp, germany_republic, tardo, lga, hifab
- **Installers:** rwedwe, lga, tardo, 'centre', 'community'
- **Extraction type:** other
- **Quantity group:** seasonal
- **Source:** lake, rainwater harvesting
- **Waterpoint type:** other

Remaining Features Frequently in Top 50 Importance Lists

- **Extraction type class:** handpump, submersible, motorpump
- **Quantity group:** enough, insufficient
- **Source:** spring, river, shallow well, machine dbh
- **Waterpoint type:** communal standpipe, hand pump, communal standpipe multiple
- **Payment:** pay per bucket, pay monthly, unknown, pay annually, pay when scheme fails
- **Permit:** True
- **Management:** vwc, wug, water board, private operator, wua
- **Water quality:** unknown, soft, salty

Future Work

- Obtaining more information about the following features could improve analysis:
  - 'Other' in the category 'extraction_type_class' (the majority of these wells are non-functional)
  - Geographic locations of funders and/or installers, e.g., Tanzania, Germany, US, Great Britain
  - Year constructed—No date recorded for significant percentage of wells; do they share some common features, e.g., location, type of well
- Performing analysis by latitude/longitude could reveal status differences due to geologic variation