Tanzania’s Water Pumps

Analysis and Interpretation

by Leticia Drasler

What if you can't have access to clean water?

- In Tanzania most of the access to clean water is through water pumps.
- According to 2015 Tanzania Water Point Mapping Data a significant portion of these water points are non-functional.
- Could we predict which ones are likely to fail?
- Overall goal is to maintain water supply and forecast repair costs by region by having predictions of which pumps will fail.

Data

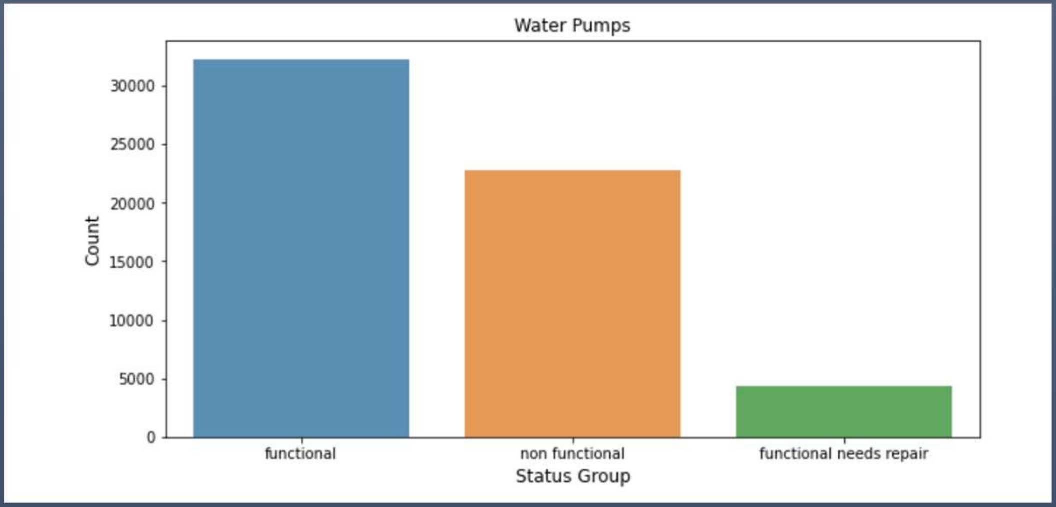
Our goal through this project is to use Machine Learning models to make predictions on the functional status of water pumps with unknown status.

The Tanzania Ministry of Water has been feeding detailed dataset, that contain about 59 thousand values in 39 variables, that include the description of each of these water points, such as:

- Location,
- Funder,
- Installer,
- Etc.

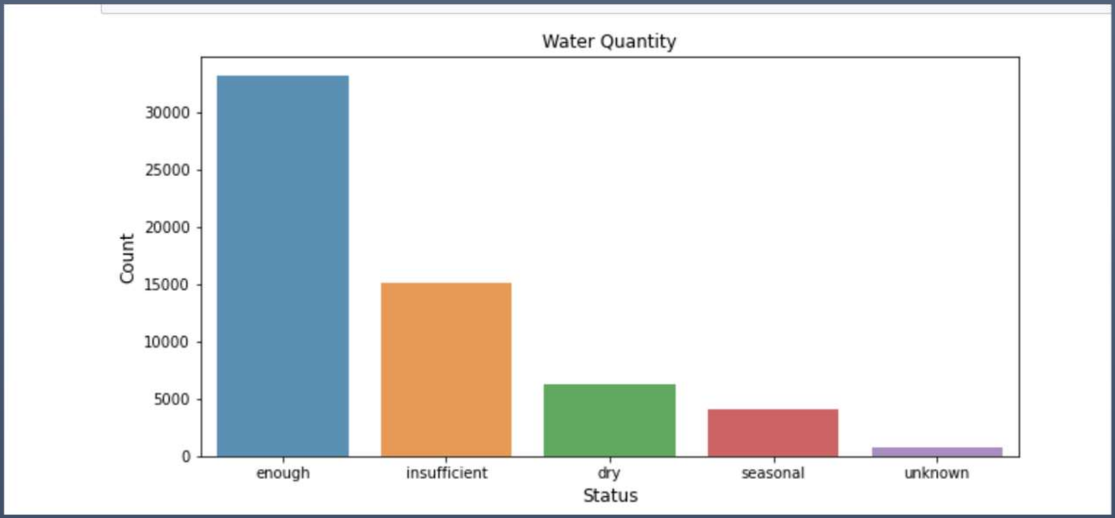
These data were provided by Driven Data.

ANALYSIS



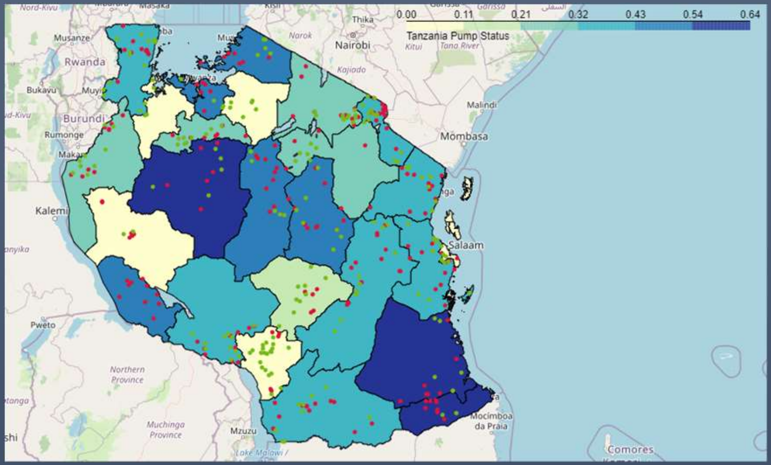
These data show us that almost half of the water pumps are non-functional

ANALYSIS

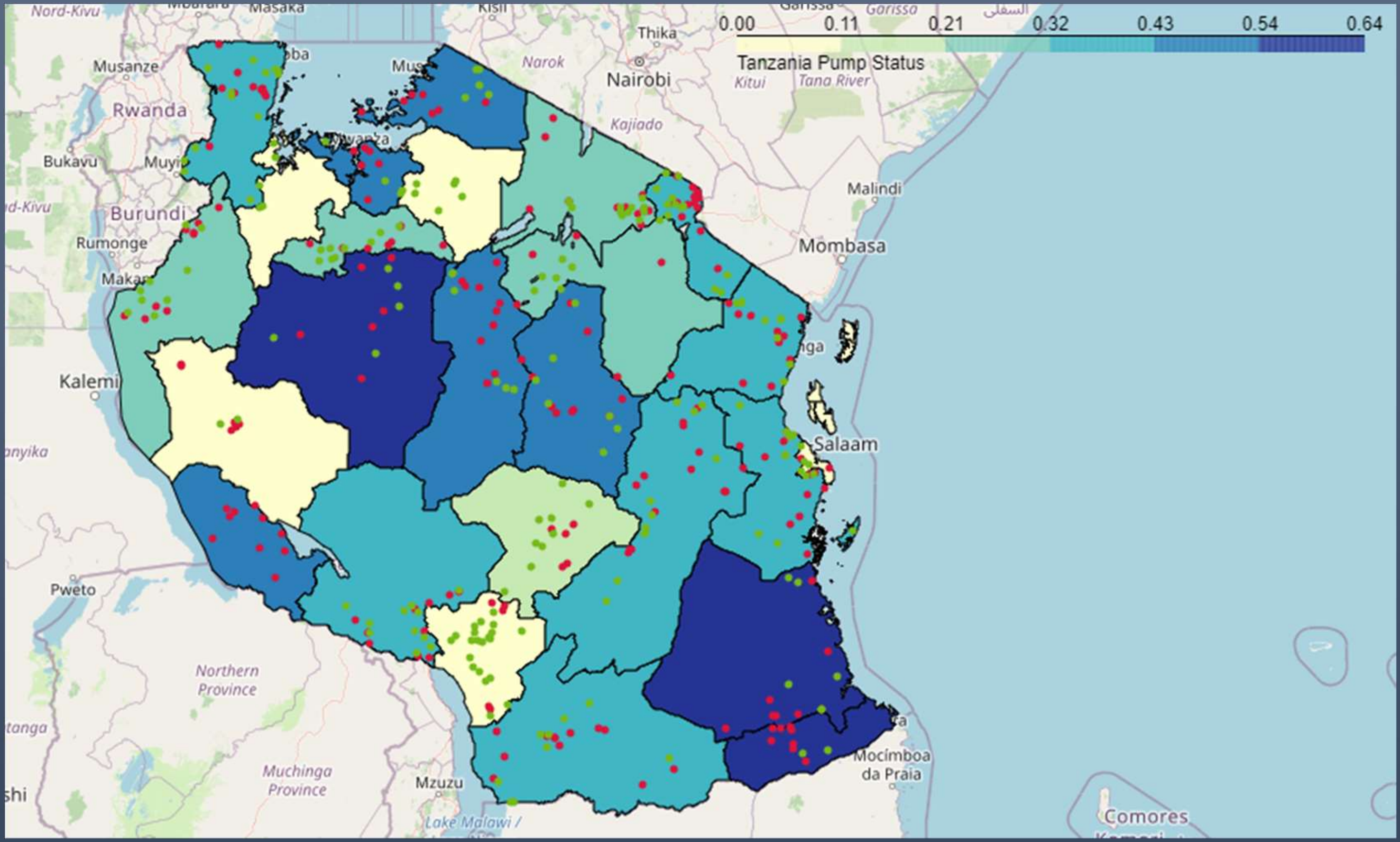


Water insufficient and dry regions also aggravates the functionally of the water pumps.

ANALYSIS



We Plot the regions for a better understanding of the functional status of water pumps based on geography. Color is a function of the proportion of non-operational pumps.



Regional map of Tanzania illustrating a sample of functional and non-functional pumps.

Modeling and Methods

For this analysis I applied multiple models to compare the results

Methods:

- Roc Curve
- Confusion Matrix

Models:

- DecisionTreeClassifier
- RandomForestClassifier
- BaggingClassifier
- Adaboost
- GradientBoost
- XGBoost

Results

Final Model

XGBoost 0.785%

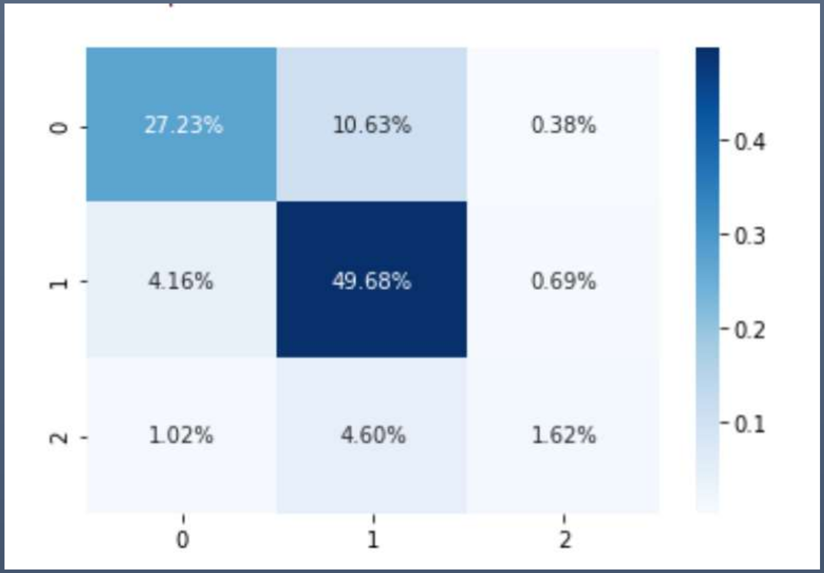
- Our initial models gave reasonable predictions but gave more errors than our final model.
- Utilizing XGBoost, we were able to predict the functional status of unknown water pumps 79% of the time.

Results

Final Model

XGBoost 0.785% Test Accuracy

- The Confusion Matrix illustrates how well our model predicts each label for the test data.
- We can see that our model classifies 10% of broken pumps as functional.
- Our model classifies 4% of functional pumps as non-functional.
-



True Label	0	1	2
0	27.23%	10.63%	0.38%
1	4.16%	49.68%	0.69%
2	1.02%	4.60%	1.62%

Predicted Label

Conclusion

- The analysis leads us to believe that the label status of the water pumps were best predicted using the XGBOOST algorithm.
- This algorithm achieved a 79% accuracy when classifying our test data.
- From these results, we feel confident that these data contain enough information to make meaningful recommendations regarding where to direct repair resources and how to best prevent interruptions in the water supply.
- We advise the state department to make accurate records of the construction year, altitude of the well, water quality, and water quantity, as these were useful parameters in our predictions.

Future Work

- Data completeness and accuracy:
  - Based on our models, we can direct the agency in how best to record data on each well. We noted construction year, altitude of the well, water quality, and water quantity as parameters to validate.
- Data transformation:
  - We will continue to look over our data and determine if other transformations such as dimensionality reduction may help to better analyze our data.
- Algorithm:
  - Inside of the machine learning community, there is continual development in the classifier algorithm. We will continue to explore these and apply them to our problem.

THANK YOU!

Leticia Drasler

Local children pump ground water in the Mara River Basin, Tanzania.

© Ana Lemos 2016 http://maraselva.fiu.edu/en/fetching-water-on-a-saturday-morning/