Pump It Up: Data Mining the Water Crisis

—

By Christa Dawson and Dariga Kokenova

Tanzania:

Population: 57 million

Lack basic access to safe water: 24 million *(Source: lifewater.org)*

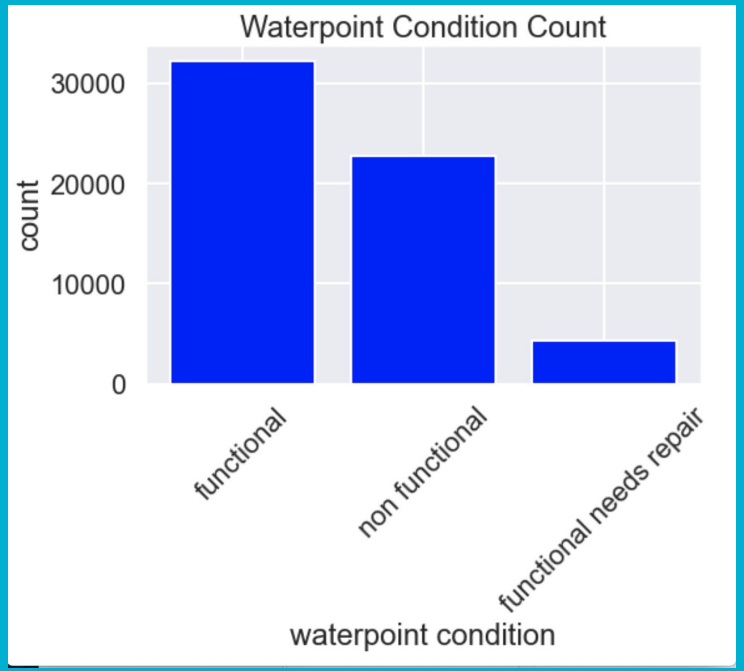
Using data from Tanzanian Ministry of Water and nonprofit organization Taarifa we built a model to predict which waterpoints are functional or not. An accurate understanding of which waterpoints are non operational can improve maintenance and enhance access to clean water across the communities.

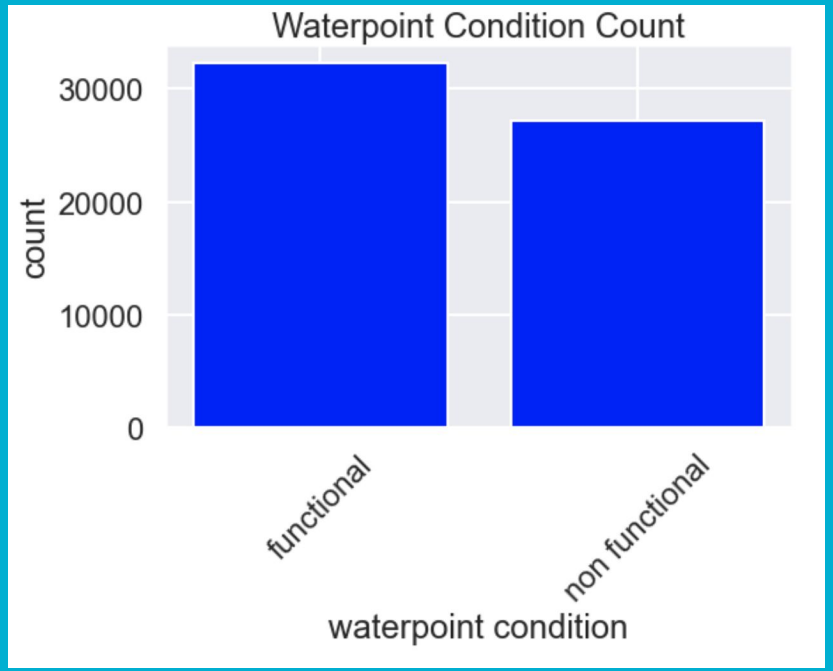
We were able to predict non functional waterpoints with 82% accuracy and 74% sensitivity.

Data overview:

- Source: Tanzanian Ministry of Water and nonprofit organization Taarifa
- Data volume: 59400 observations by 39 columns
- Range of Years for Data Collection: 2002 - 2013
- Data Cleaning Process:
  - Addressed class imbalance of target variable
  - Remapped installer column. If the count<200 or NaN, map to category “other”
  - Filled NaN for public meeting and permit with False
  - Addressed construction year=0 :
    - First remapped based on region mean
    - For regions missing construction year altogether, remapped with overall construction_year mean
    - Fixed observations with construction year>data collection year
- Data Engineering: added column age equal to data collection year - construction year

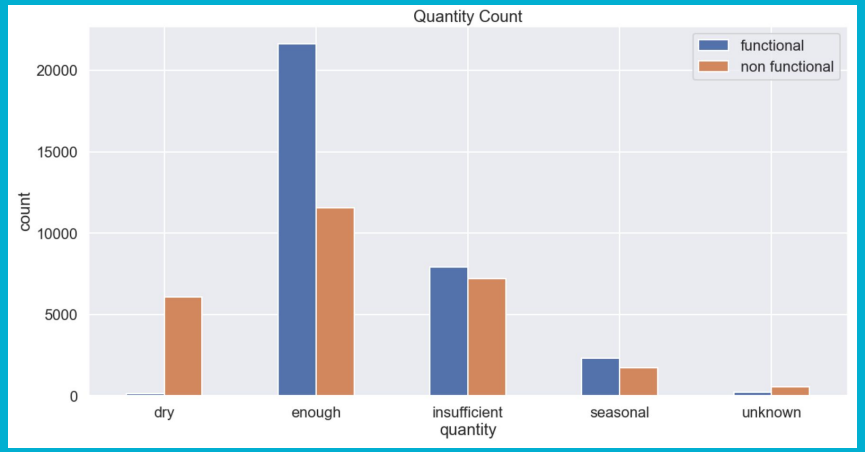
Waterpoint Condition Count: Class Imbalance

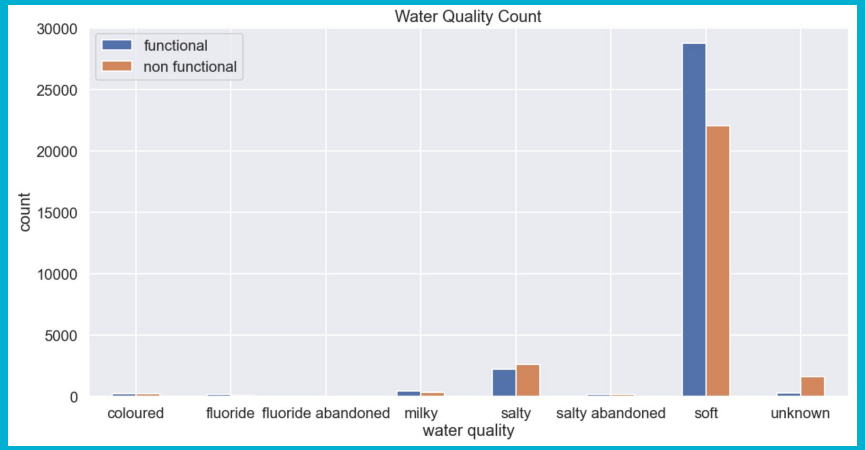




Null accuracy (if we always predicted majority class) = 32359/(32359+27141) = 54.38%

Waterpoint status by quantity and quality

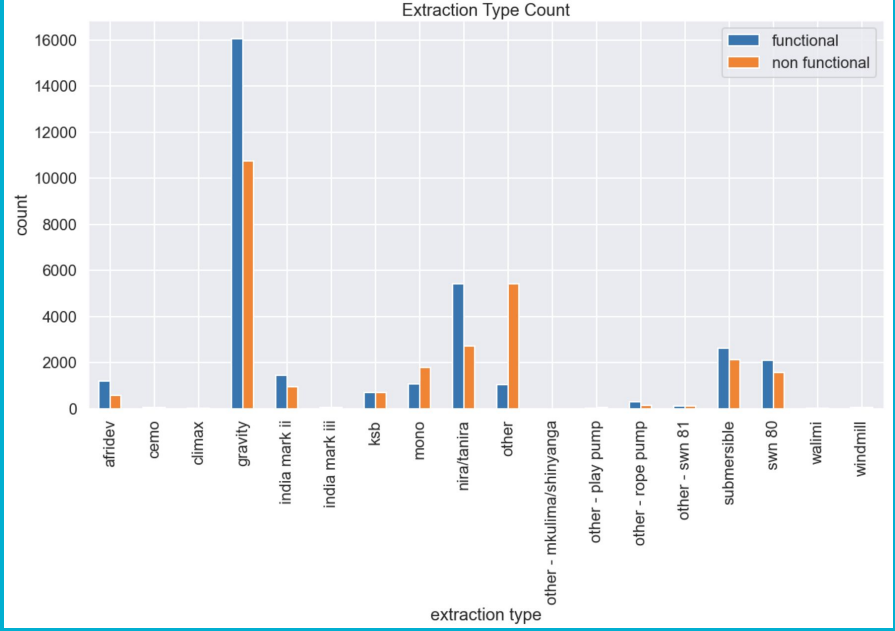


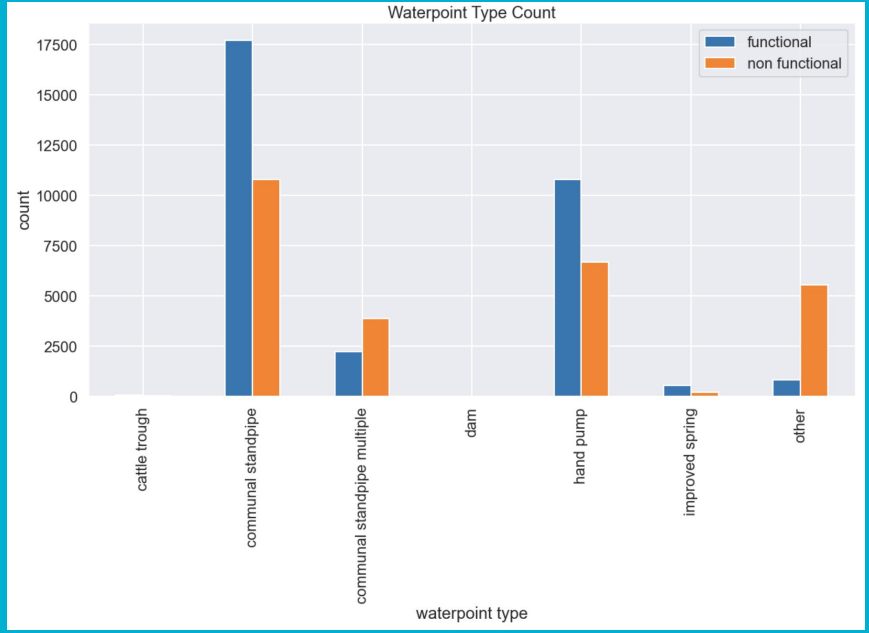


We expect more functional waterpoints when quantity is enough and more non functional when quantity is dry.

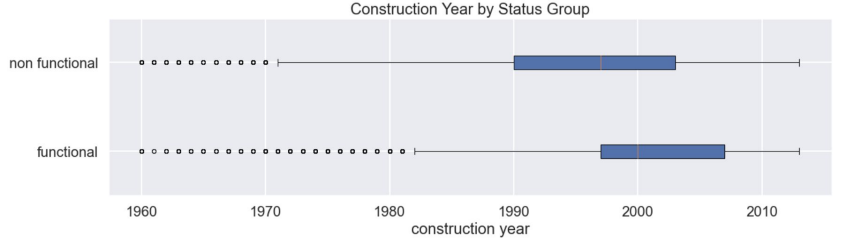
We expect more functional waterpoints when water quality is soft, and less difference between other categories.

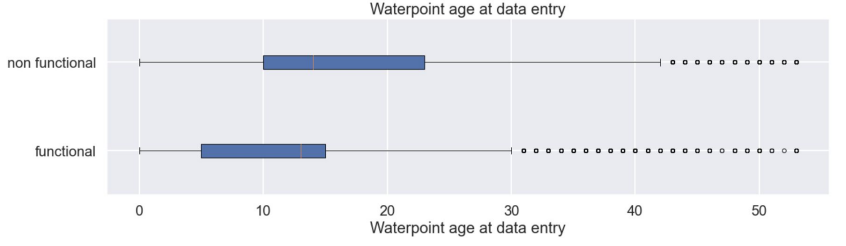
Waterpoint status by extraction type and waterpoint type





Waterpoint status by construction year and age





Older waterpoints are more likely to be non functional.

Construction year is highly correlated with age, thus decided to remove it from the model.

Classification Evaluation (Confusion matrix metrics):

⭐ **Accuracy**: how often are we correct at predicting functional and non functional waterpoints?

Accuracy = (TP + TN) / (all predictions)

Precision: when we predict the waterpoint to be non functional, how often is that prediction correct?

Precision = TP / (TP + FP)

⭐ **Recall (Sensitivity)**: what proportion of truly non functional waterpoints was identified correctly?

Recall (Sensitivity) = TP / (TP + FN)

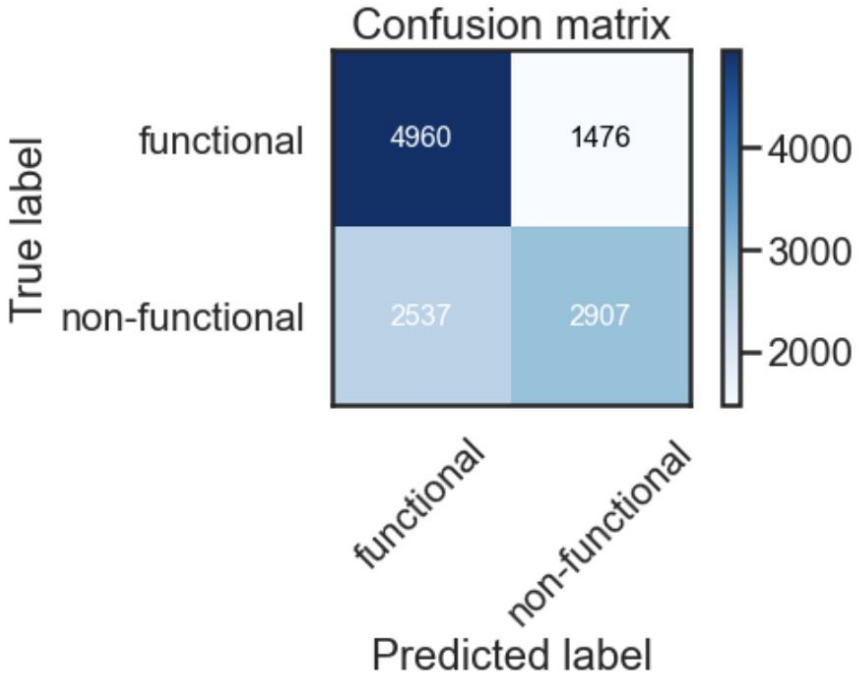
F1 score: harmonic average of precision and recall metrics.

F1 score = (2 * Precision * Recall) / (Precision + Recall)

First simple model: Logistic Regression

Test Accuracy score: 66.22%
Train Accuracy score: 66.71%

Test Recall (Sensitivity) score: 53.40%
Train Recall (Sensitivity) score: 53.61%



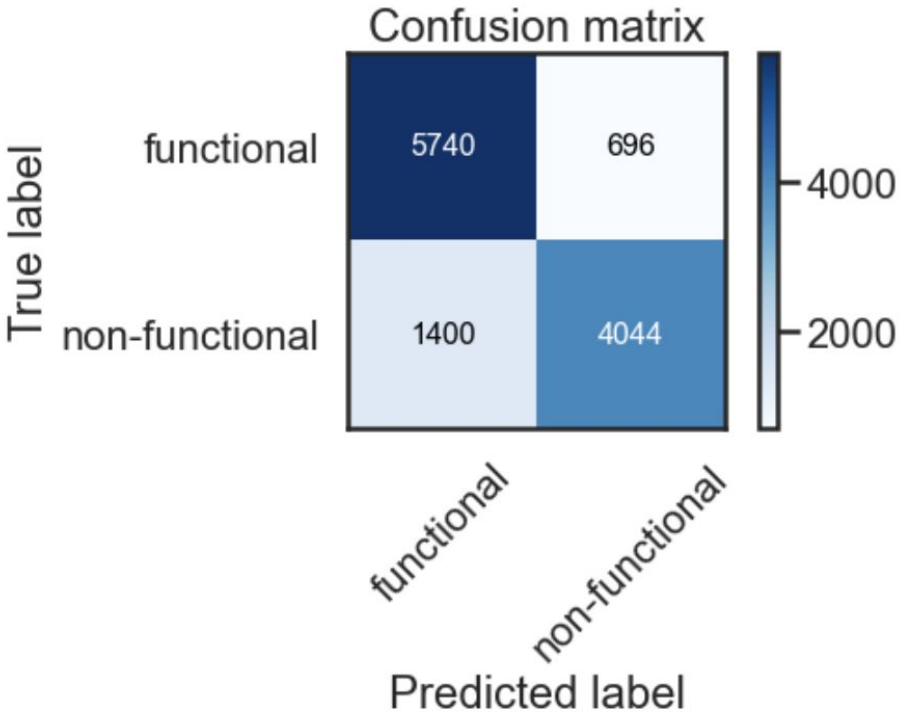
Models used:

- Logistic Regression
- K Nearest Neighbors with search for optimal hyperparameter k
- Decision Tree with grid search
- Random Forest with grid search
- AdaBoost
- Gradient Boost

Final model: Random Forest (max_depth=20)

Test Accuracy score: 82.36%
Train Accuracy score: 93.10%

Test Recall (Sensitivity) score: 74.28%
Train Recall (Sensitivity) score: 87.48%



Next steps:

- Do not combine “functional needs repair” category with “non functional” category and model the target variable using resampling techniques
- Look into observations with population 0. 21,381 observations in total
- Add recursive feature elimination
- Add more parameters to existing grid searches
- Add grid search to AdaBoost and Gradient Boost

Questions?

GitHub: https://github.com/dawsonc96/Tanzanian-Water-Well

Contact Information:

Christa Dawson https://www.linkedin.com/in/christa-dawson-a955bb205/

Dariga Kokenova https://www.linkedin.com/in/darigakokenova/