Identifying Non-functioning Wells in Tanzania

Modeling and Analysis by Jack Maninx

Summary

- WaterTanzania needs a model that can predict the functional status ('Functioning' or 'Non-Functioning') of a well
- Conduct outreach to responsible entities to be notified when a well is in need of repair
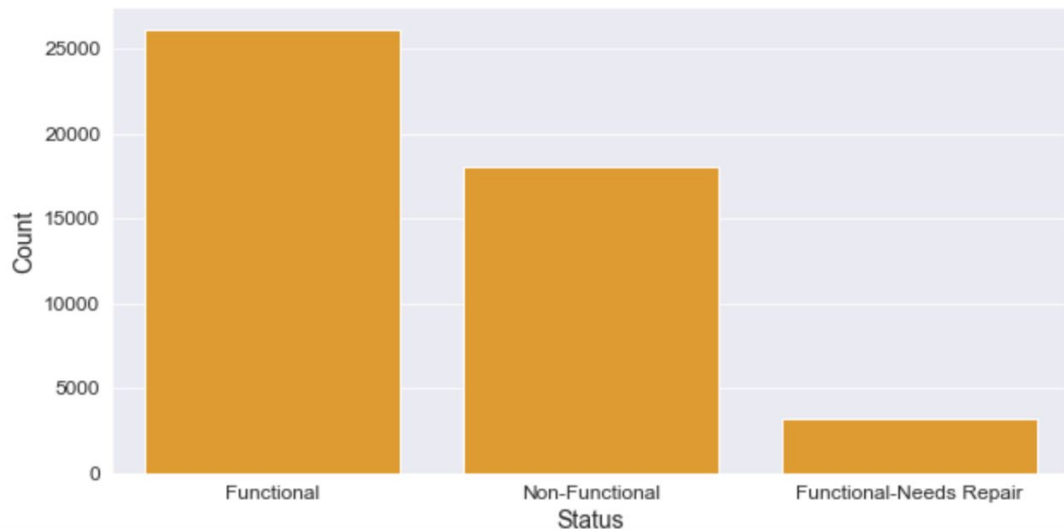
Legend

- *NF* = Non-Functional
- *F* = Functional
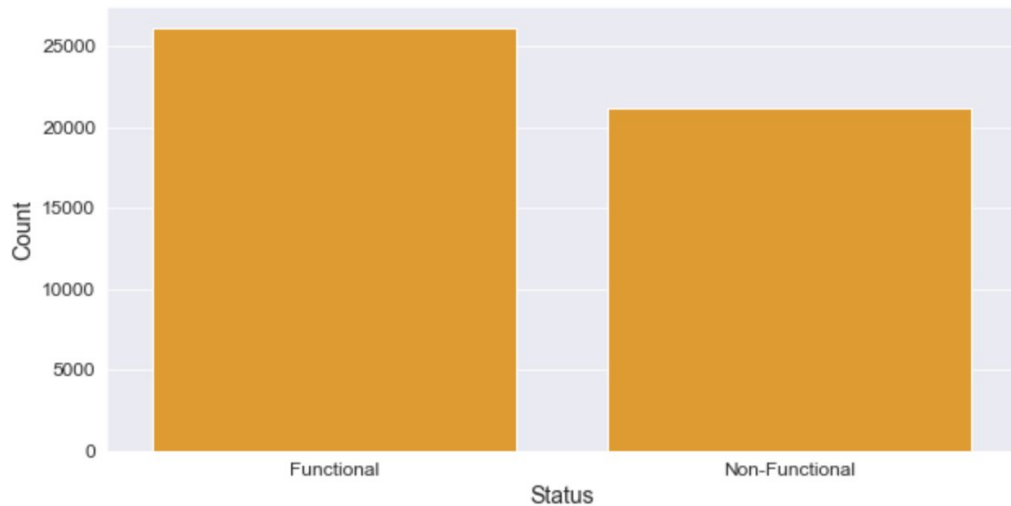- *FR* = Functional - Needs Repair

Data & Methods

- Data provided by TAARIFA & DrivenData
- 47,304 wells analyzed
- Data collected 2011 - 2013
- Models Used:
  - Logistic Regression
  - Decision Tree
  - RandomForest
  - XGboost

Assumptions - Class Imbalance

- Class imbalances pose a massive obstacle for modeling
- Early model performance was poor for *FR*
- False reports of wells being *FR* when they are *NF*
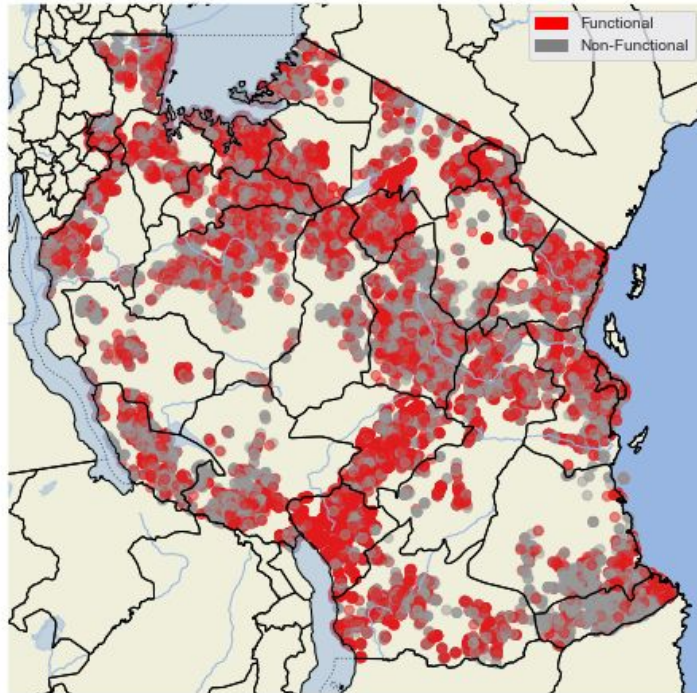 - Little urgency

Assumptions - Class Imbalance

- *NF* and *FR* both require repair
- Aggregating *NF* and *FR* to offset class imbalance
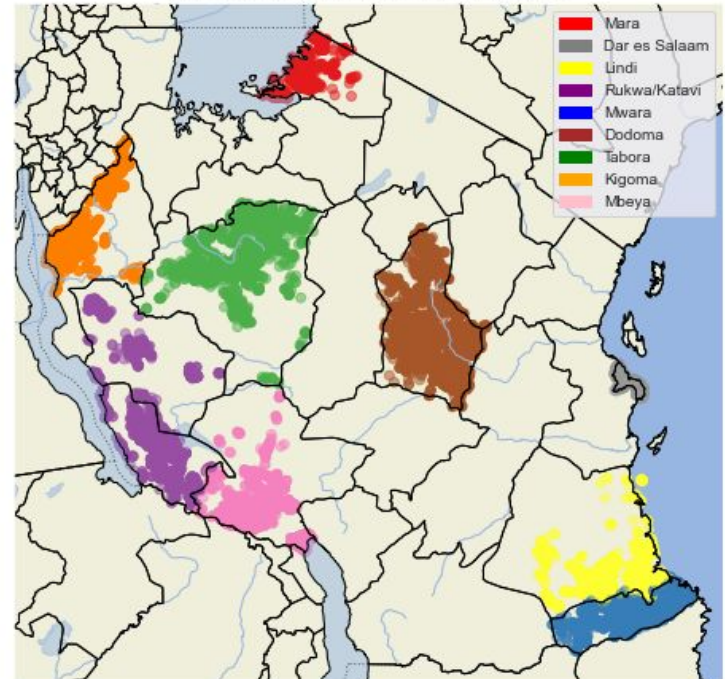
Assumptions - Class Imbalance

- Worst Case:
  - *NF* well being labeled as *F*
  - *False Negative*
- Better Case:
  - *F* well being labeled as *NF*
  - *False Positive*
- Best Case:
  - *F* and *NF* wells being labeled correctly
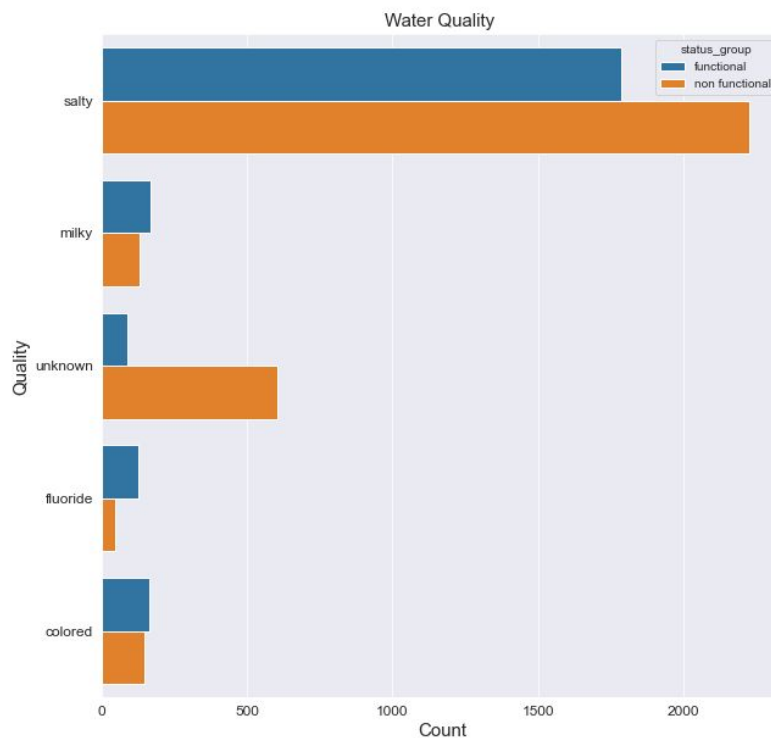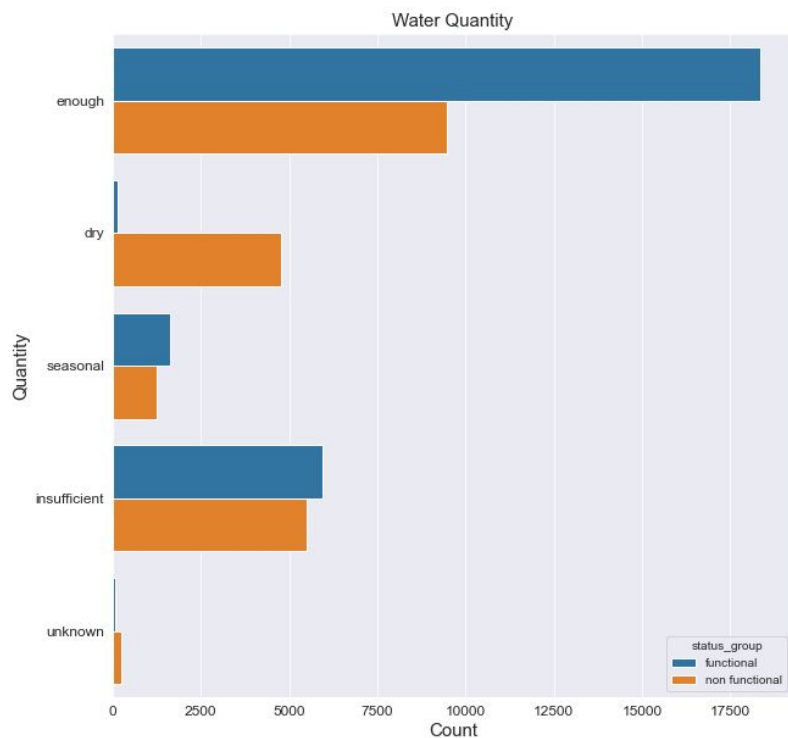 - *True Positive / True Negative (respectively)*
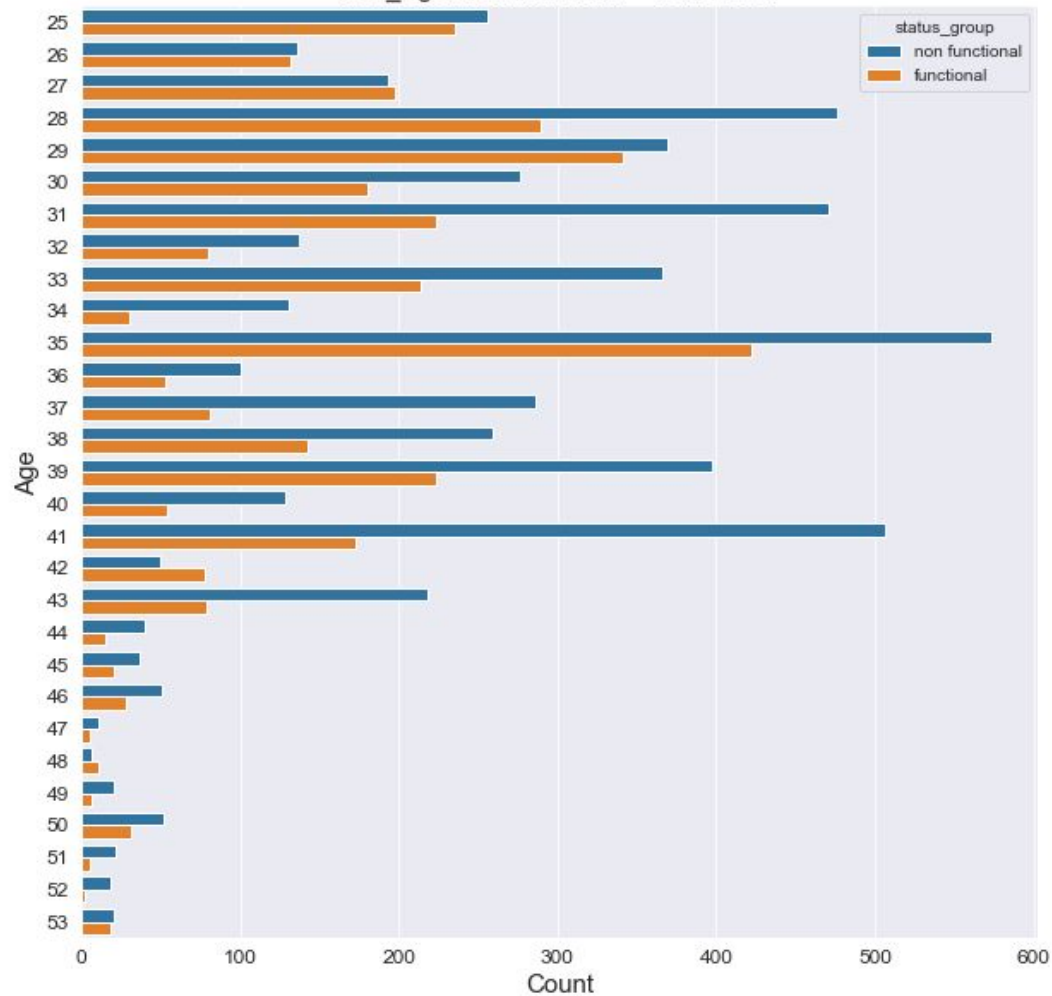
Data & Methods

Functional vs Non-functional Map
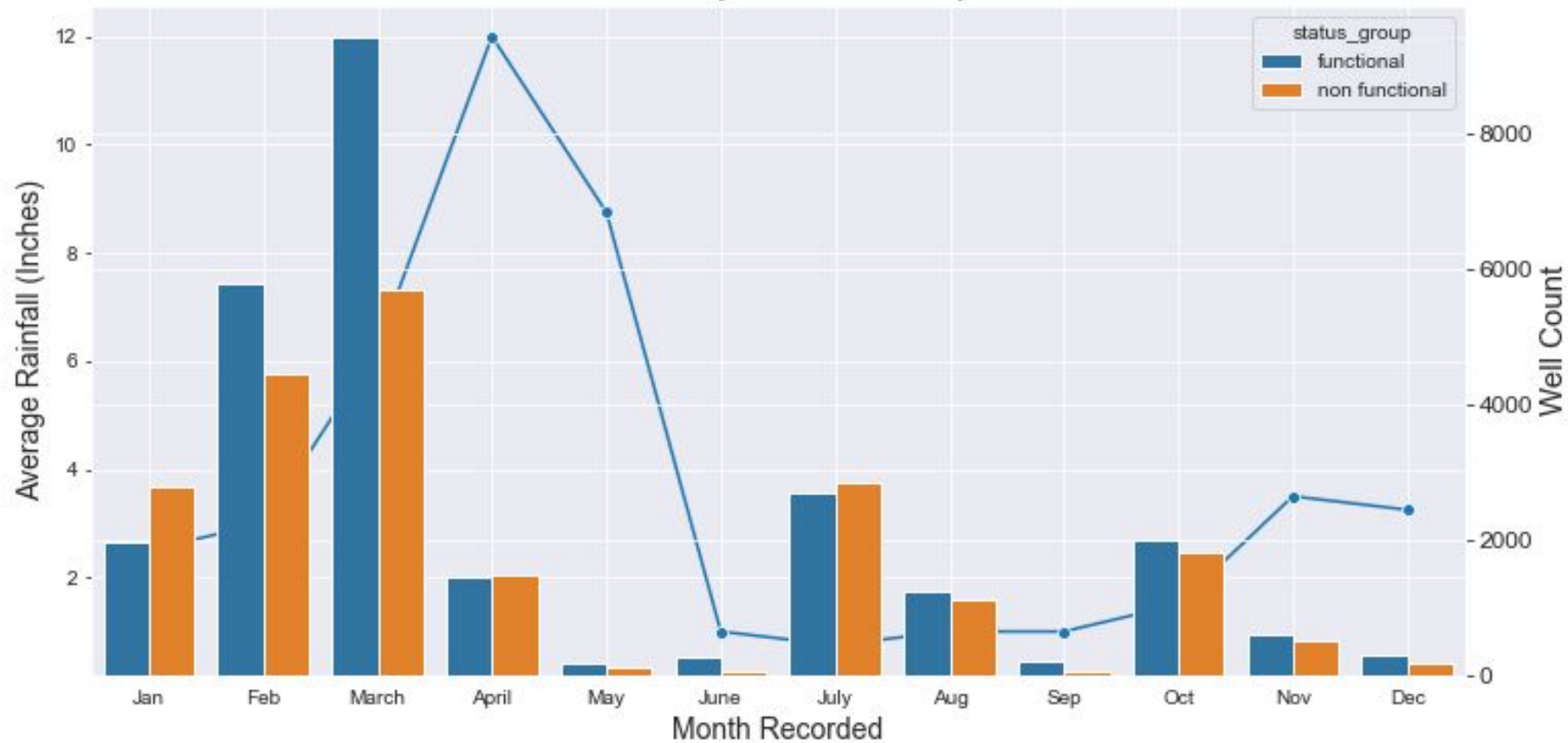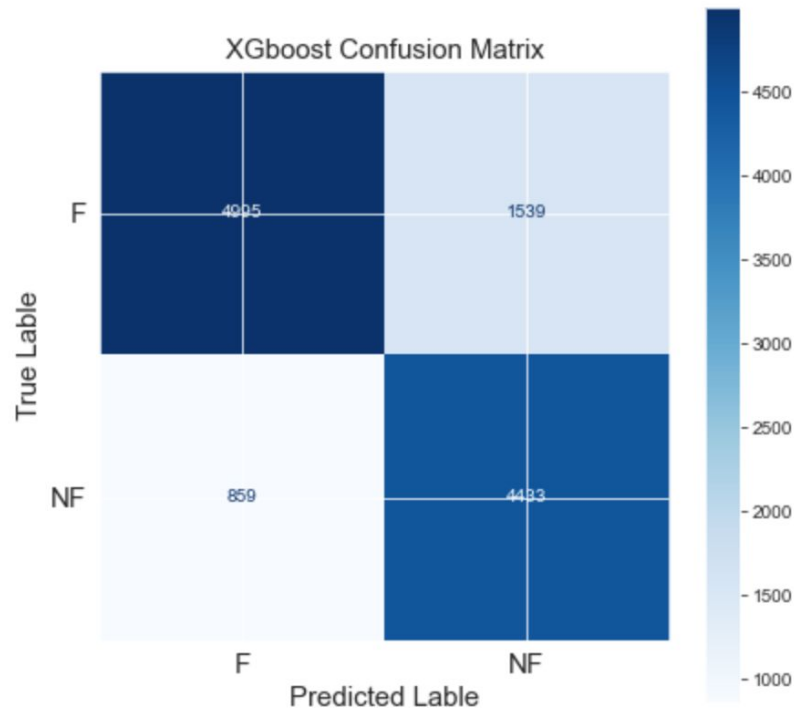
Regions: Non-functional > Functional

Data & Methods

Well_Age: Non-functional > Functional

Wells Recorded by Month & Monthly Rainfall

Results

XGboost Confusion Matrix

- **F** Recall : 76%
- **NF** Recall : 84%

Recommendations

Monitoring

- Regions in which **NF** wells outnumber **F** should be monitored with heightened attention:
  - *Mara, Mtwara, Tabora, Rukwa, Kigoma, Lindi, Dodoma, Mbeya, Dar es Salaam*
- Monitoring efforts should increase during months in which **NF** wells outnumber **F**:
  - January, April, July

Alerts

- Send warning when a well ages over 24 years
  - Suggest more frequent checks
- Send warning anytime a well's water quantity is reported as 'insufficient' and/or 'dry'
  - Suggest checking
- Send warning in June when dry season begins
  - Suggest more frequent checks through October

Going Forward

- Conduct PCA during initial data cleaning in order to reduce dimensionality in a more calculated manner
- Advocate for additional metrics to be used to describe "Water Quality"
- Advocate for higher recording rates during dry season

Thank You!

Jack Mannix

Email: jmannix15@gmail.com

Github: Hugh_Jackman_nix