Tanzanian Water Crisis:

Predicting Functionality of Water Wells in Tanzania Using Machine Learning

By: Orin Conn

Table of Contents

- The Problem & Data Utilized

- OSEMN Modeling Process

- Best Machine Learning Model

- Most Important Features of Functionality

- Conclusions & Recommendations

The Problem:

- ⅓ of Tanzania is arid or semi-arid (desert)
- No access to clean water
- Pollution contaminates groundwater
- Pumps don't always work.

Data Utilized:

- DrivenData dataset
 - Tanzanian Ministry of Water
  - Taarifa

OSEMN Process

- **Obtained Compiled Data from DrivenData**
- **Scrub missing values & duplicate variables**
  - **Converted categorical data to numeric**
- **Explored relationship to functionality**
  - **Water Quality**
  - **Pump Types**
  - **Location, etc.**
- **Modeling**
  - **Decision Tree**
  - **KNN**
  - **Random Forest**
  - **XGBoost**
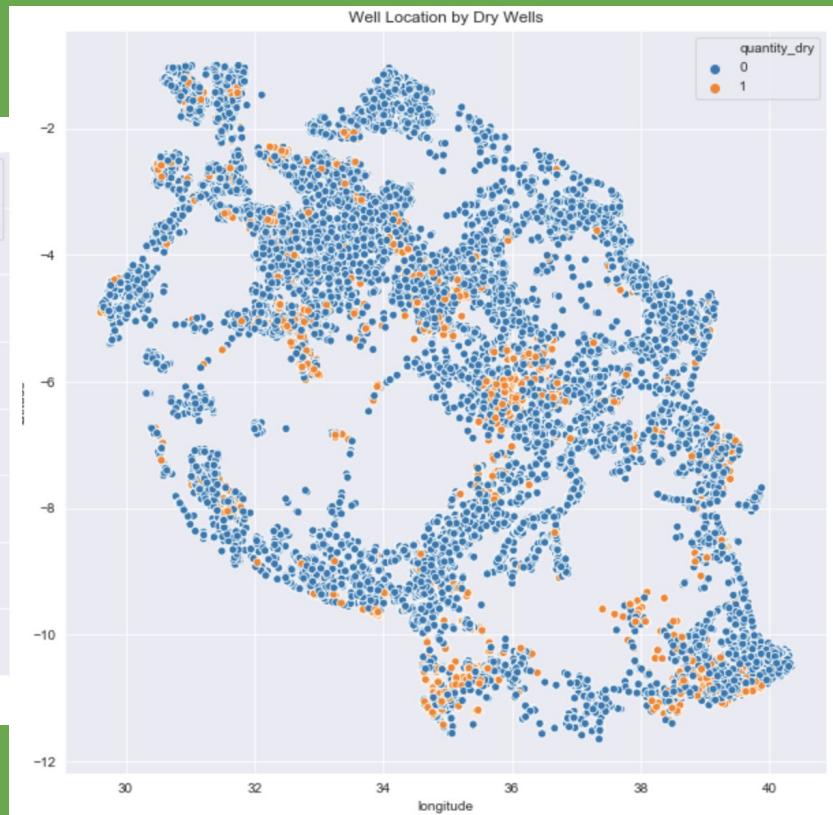- **iNterpreted best results!**
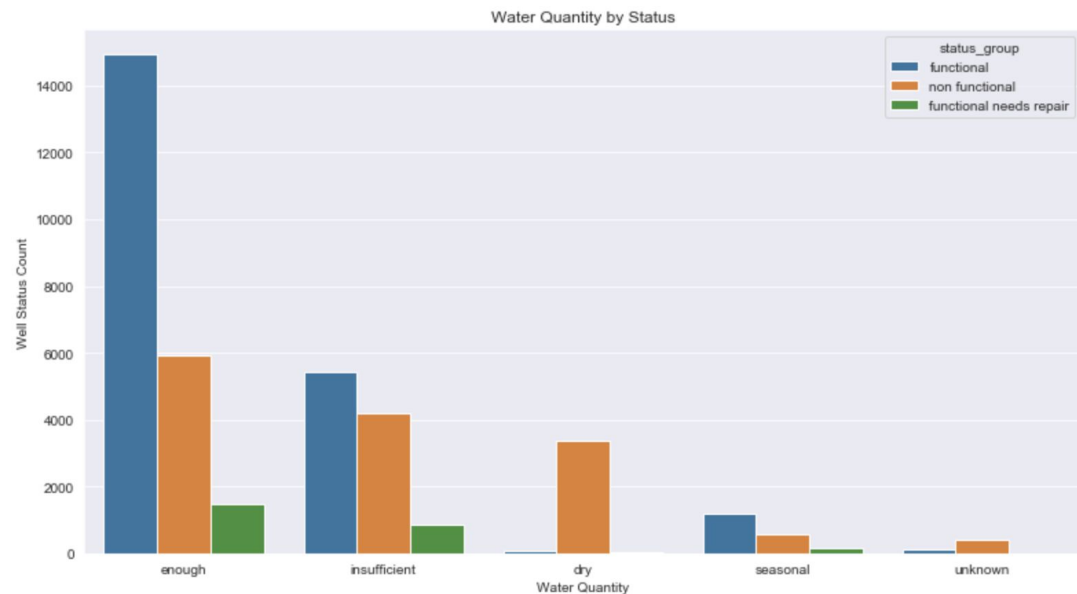
Best Performer: Random Forest

Highest Accuracy: **0.8095**
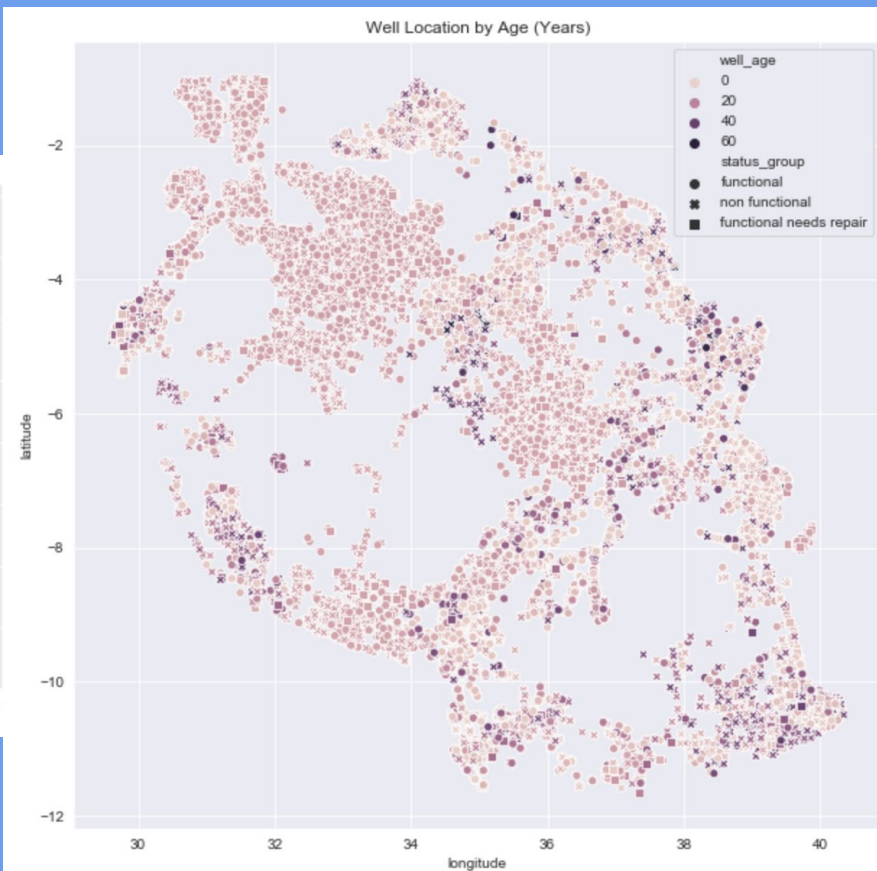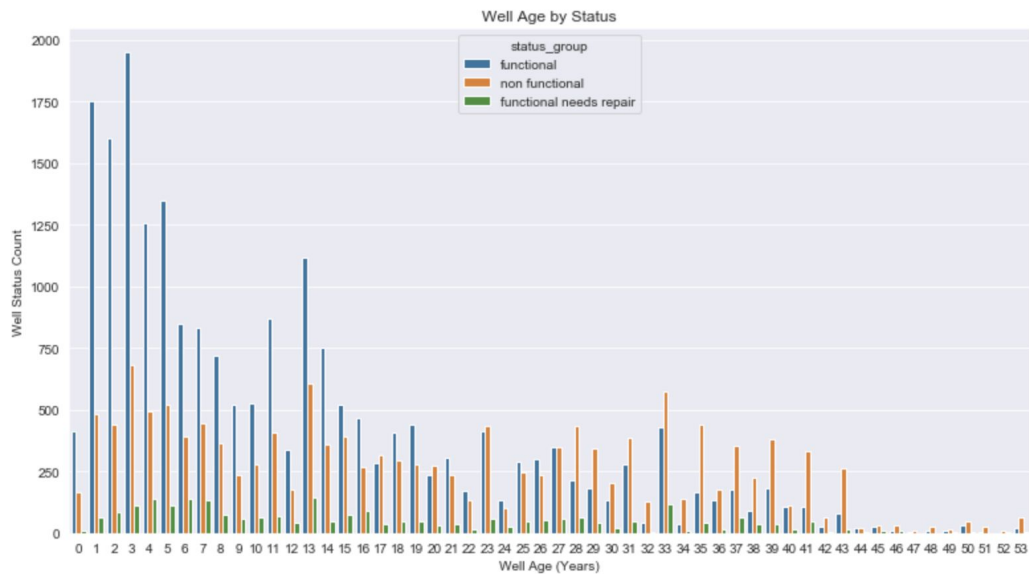
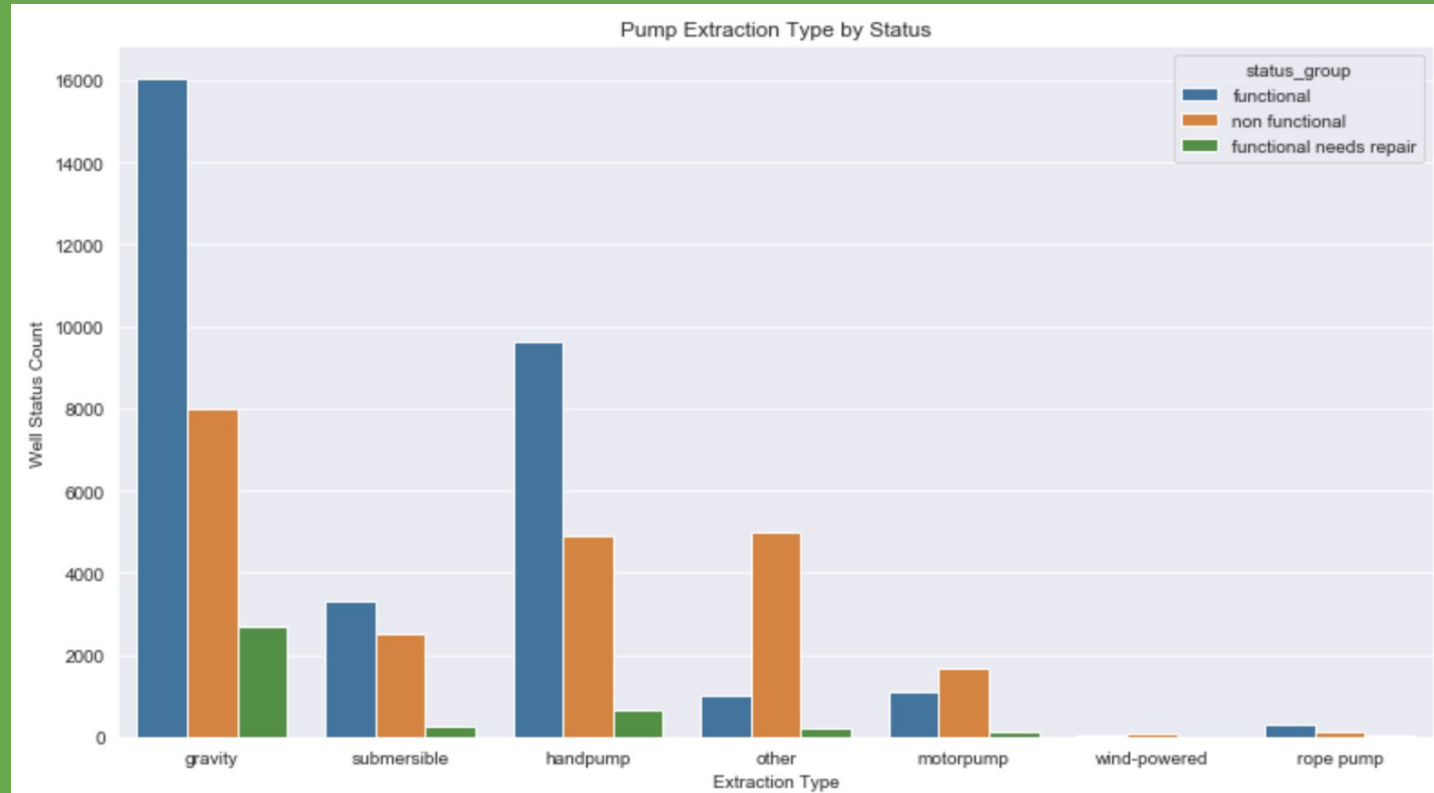n_estimators = 200,

min_samples_split=8
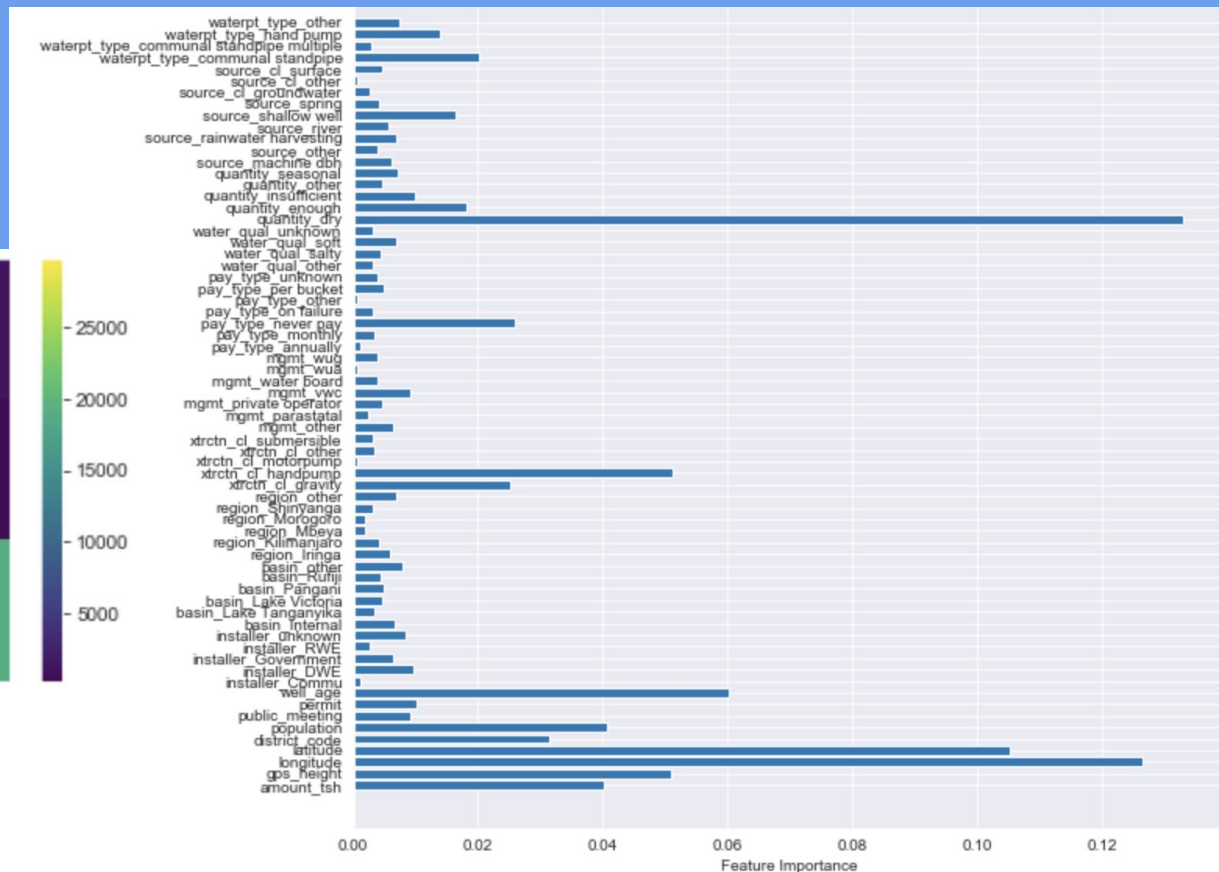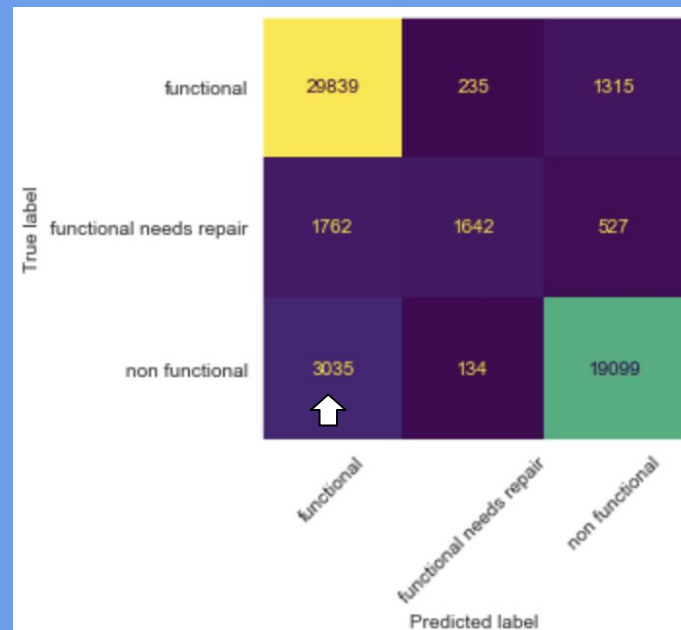
Water Quantity of Well: Dry = Non-Functional

Age of the Well Pump: Date of Data Recorded - Construction Year

Pump Extraction Type: gravity, handpump - mostly functional

Pump Extraction Type by Status

Conclusions

- ⬇ water quantity = ⬇ functionality
  - Dry wells are almost always not working
- ⬆ well age = ⬇ functionality
  - Older pumps should be refurbished
- Gravity pumps and handpumps are less prone to breaking
  - Only install these if possible

Future Recommendations

- Limit the redundancy of categorical data
- More time to explore amount of use per well, possibly population
- Exploration into cost associated with well type

Thank you!!