Tanzania Water Wells

Predicting Well Conditions to Optimize Resource Allocation

PATRICIA FULGO

Business Challenge & Stakeholder Needs

- Our project tackles the challenge of identifying which water wells in Tanzania need repair, so that organizations like WaterAid and government agencies can better direct their resources

•**Business Problem:**

•Tanzania has over 57 million people, but many water wells are non-functional or require repair.

•This leads to inefficient resource allocation and impacts public health.

•**Stakeholder Focus:**

•**NGOs (e.g., WaterAid):** Need to quickly locate and repair failing wells.

•**Government:** Wants to understand failure patterns to guide future well construction.

Data Overview

•**Data Sources:**

•**Features Data:**
Contains information about each well (e.g., pump type, installation date, location, funder, etc.).
•**Labels Data:**
•Contains well condition information (functional, non-functional, needs repair).
•**Dataset Size:**
•Approximately 59,400 records.

•**Key Attributes:**
•**Numeric:** Amount of water available, altitude, GPS coordinates, population.

•**Categorical:** Funder, installer, region, extraction type, etc.

•**Binary:** Whether a public meeting was held or a permit exists.

Data Preparation & Methodology

Data Cleaning:
•Removed non-predictive fields (e.g., IDs, recording dates) and corrected missing values.
•Converted dates into a new feature: **Well Age** (derived from installation date).

Feature Engineering:
•Consolidated rare categories (e.g., funders with few occurrences labeled as "Other").
•Re-coded target variable into numeric codes (0 for non-functional, 1 for functional, 2
•for needs repair).

"We built our model using a two-step process: a baseline model using Logistic Regression for interpretability and an advanced model using a Random Forest for improved performance.

Modeling Approach

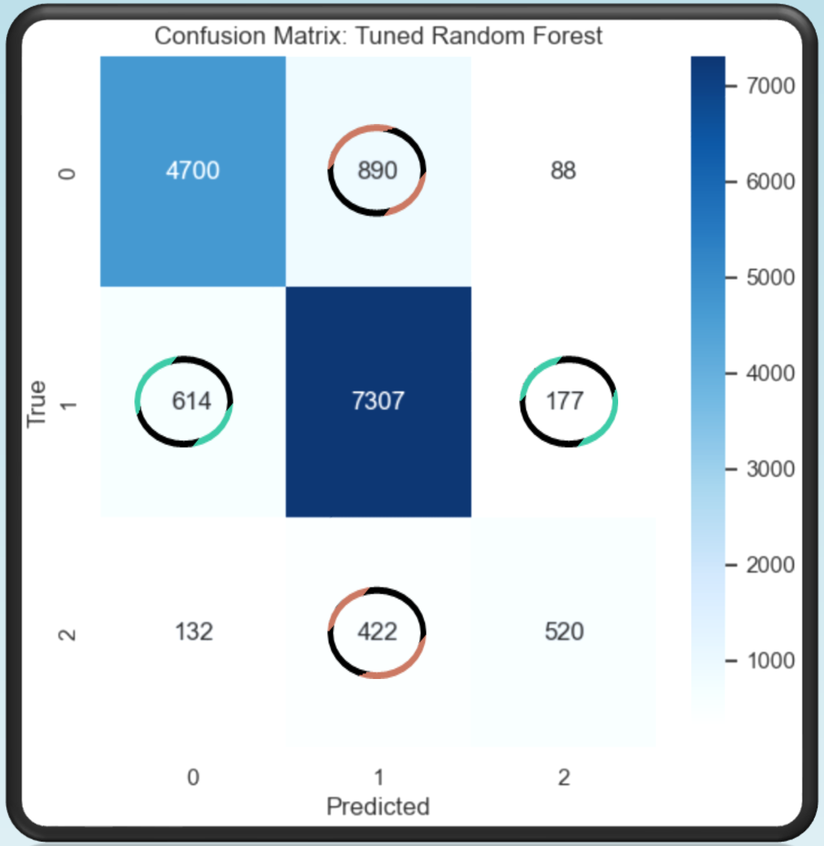
Baseline Model:
•**Technique:** Logistic Regression
•**Purpose:** Set an initial performance benchmark that is simple and interpretable.

Advanced Model:
•**Technique:** Random Forest with hyperparameter tuning (via RandomizedSearchCV)
•**Purpose:** Improve prediction accuracy and capture complex patterns.

"Our iterative modeling approach allows us to understand trade-offs and improve performance step by step."

Results & Evaluation

Waste resources
People in need



•**Confusion Matrix:** Visualizes model errors to show which well conditions are correctly or incorrectly classified.

•**Model Comparison:**
•**Baseline Model:** Achieved around 75% testing accuracy.
•**Advanced Model:** Improved overall performance with better-balanced class predictions.

Business Implications

Actionable Insights:

•**For NGOs:** Use the model to prioritize repairs for wells flagged as non-functional or needing repair.

•**For Government:** Use the analysis to understand key factors (like well age or pump type) that affect well performance, guiding future well construction.

"This predictive model empowers decision-makers to allocate resources more effectively, ultimately improving water access and public health outcomes in Tanzania."