Predicting Water Pump Functionality in Tanzania

By [Achieng Otieno]

Addressing Tanzania's Water Challenge

Tanzania faces a critical challenge in providing reliable access to clean water, with many water points serviced by pumps falling into disrepair. This project aims to harness the power of machine learning to proactively identify at-risk water pumps.

The Problem

Many water pumps are non-functional or in need of repair, hindering access to clean water.

The Goal

Develop an ML model to predict pump status: functional, non-functional, or functional but needs repair.

Understanding the Data

Our analysis is built upon a comprehensive dataset detailing over 59,000 water wells across Tanzania, enriched with 40 distinct features.

Dataset Size	Over 59,000 water well records with 40 features.
Feature Types	Numerical (e.g., amount_tsh, gps_height), Categorical (e.g., funder, basin), and Temporal (date_recorded, construction_year).
Target Variable	"status_group": functional, non-functional, functional needs repair. **Crucially, this variable is imbalanced.**

The imbalance in the target variable, particularly the smaller count of "functional needs repair" wells, was a key consideration throughout the modeling process.

Preparing Data for Prediction

Rigorous data preprocessing was essential to ensure the quality and relevance of features for the predictive models.

01

Handling Missing Values

Addressed gaps in critical features like 'funder' and 'installer' to create a complete dataset.

02

Feature Engineering

Derived meaningful metrics, such as 'age_of_well' from 'date_recorded' and 'construction_year', to enhance model understanding.

03

Encoding Categorical Data

Transformed categorical variables into numerical formats using One-Hot/Label Encoding for machine learning compatibility.

Modeling the Challenge

We explored a variety of classification models and techniques to tackle the prediction problem, especially addressing the inherent data imbalance.

Models Explored

- Logistic Regression
- Random Forest
- XGBoost

Addressing Imbalance

SMOTE (Synthetic Minority Over-sampling Technique) was applied to oversample the "functional needs repair" class, preventing model bias.

Hyperparameter Tuning

- Randomized Search using the best parameters.

Key Challenge

> "Accurately predicting 'functional needs repair' was challenging due to its limited sample size and subtle differences from other statuses."

Model Evaluation: XGBoost Performance

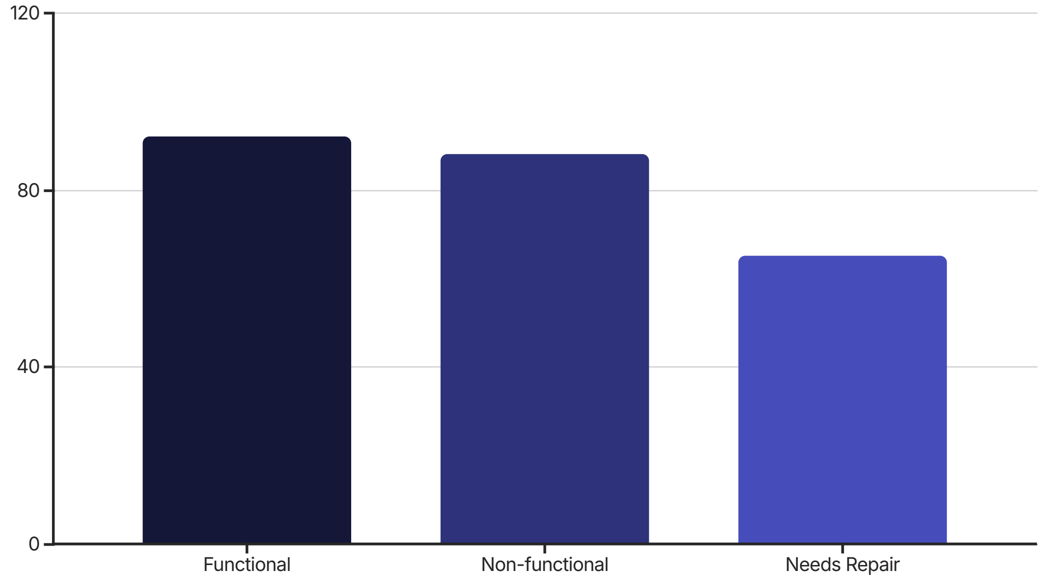
The XGBoost Classifier emerged as our top performer. While exhibiting high overall accuracy, its specific strengths and weaknesses became evident upon closer inspection.

Final Model Choice

XGBoost Classifier was selected for its superior predictive capabilities compared to other models.

Key Metrics

- Accuracy Score
- Classification Report (Precision, Recall, F1-Score)
- Confusion Matrix



The model excelled at identifying 'functional' and 'non-functional' pumps, but struggled with 'functional needs repair'.

Optimizing for Reliability: Hyperparameter Tuning

To extract the maximum performance from our XGBoost model and ensure its robustness in real-world deployment, extensive hyperparameter tuning was conducted.

Purpose of Tuning

Fine-tune model parameters for peak performance and better generalization to unseen data.

Methodology

Utilized a **Randomized Search** approach to efficiently explore a broad spectrum of parameter combinations.

Outcome & Recommendation

The tuned model offers a more reliable and realistic performance measure, making it ideal for deployment.

This optimization ensures that the model is not just accurate on historical data, but also dependable for future predictions.

Recommendations & Future Work

While the current model provides valuable insights, its predictive power, especially for pumps needing repair, can be significantly enhanced with targeted data collection.

Enhance Data Collection

To improve accuracy, particularly for the 'needs repair' category, future data efforts should focus on enriching the dataset.

Invaluable New Features

- **Last service date:** Provides critical temporal context.
- **Types of repairs:** Helps classify specific failure modes.
- **Parts replaced/repaired:** Offers granular detail on pump health.

These features would provide stronger signals for the model, enabling it to better distinguish between fully functional, failing, and failed pumps.

Key Takeaways

Meaningful Impact

Our model can significantly aid in the maintenance and distribution of clean water in Tanzania.

Data-Driven Decisions

Leveraging ML helps prioritize interventions and optimize resource allocation for water infrastructure.

Continuous Improvement

Future data enrichment, especially on repair history, will unlock higher predictive accuracy.