Transforming Water Security: Leveraging Data to Improve Well Reliability in Tanzania

...

Presented by: Huldah Chepkoech Rotich

16th September 2025

The Challenge

- Over 30–40% of rural water wells in Tanzania are non-functional at any time.
- Communities lose access to clean water → health risks, lost productivity, unmet SDG 6 goals.
- Current monitoring is reactive: wells are only fixed once they break.

Key Question: Can we predict failing wells before they break?

Problem Statement

- How can we predict the functionality status of wells?
- How can we detect patterns that influence well longevity and reliability?

Project Objective

- Predict whether a well is functional, needs repair, or non-functional.
- Identify geospatial and operational patterns linked to failures.
- Provide actionable insights for government, NGOs, and donors to:
  - Prioritize repairs
  - Improve design of new wells
  - Reduce maintenance costs

Data Overview

Data Source: Tanzania Water Wells Dataset

Dataset Size: 59,000+ records

Features included:

- Technical (e.g., pump type, construction year, water quality)
- Geographic (latitude, longitude, elevation, region)
- Operational (funder, management type, community meetings)

Target: Well status (functional / needs repair / non-functional).

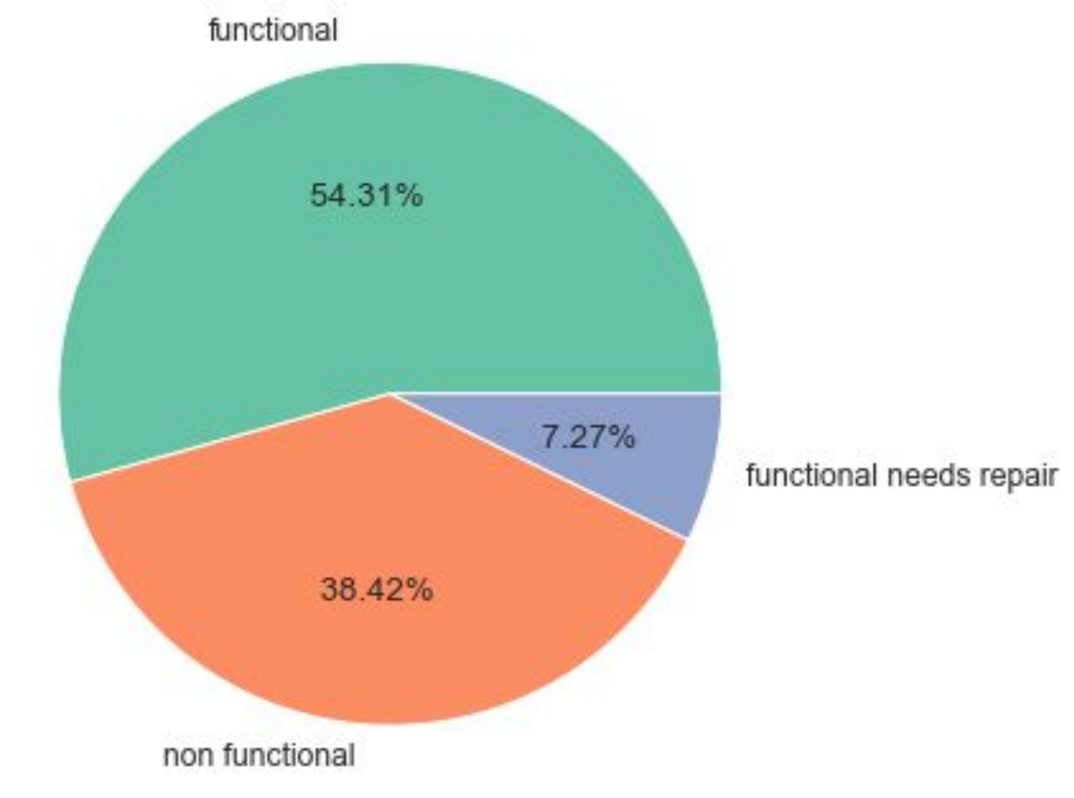
Data Preparation

Steps Taken:

- Removed nulls and irrelevant columns
- Corrected data types
- Standardised the categorical data
- Feature Engineering

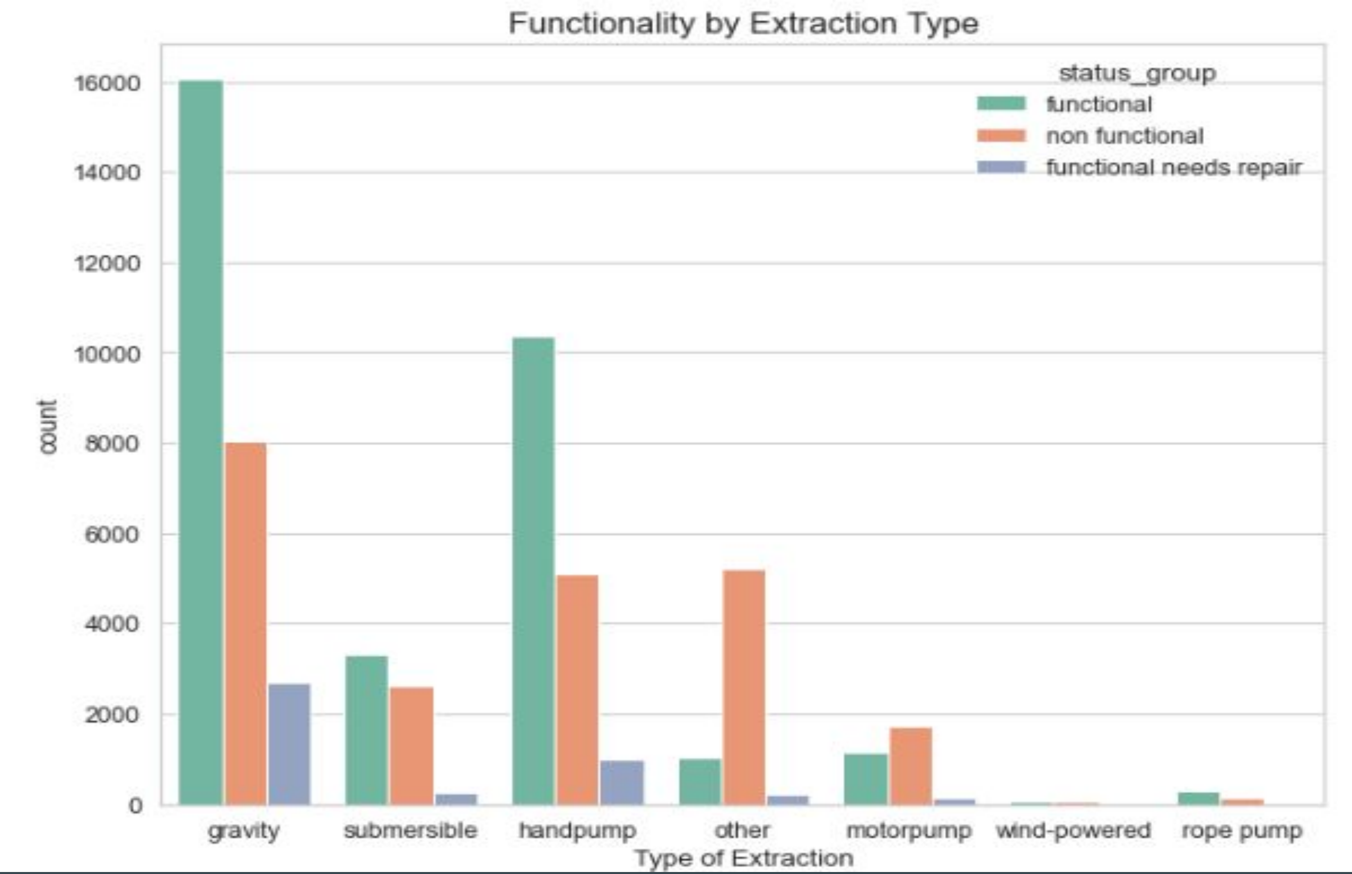
Exploratory Data Analysis

Distribution of Well Functionality

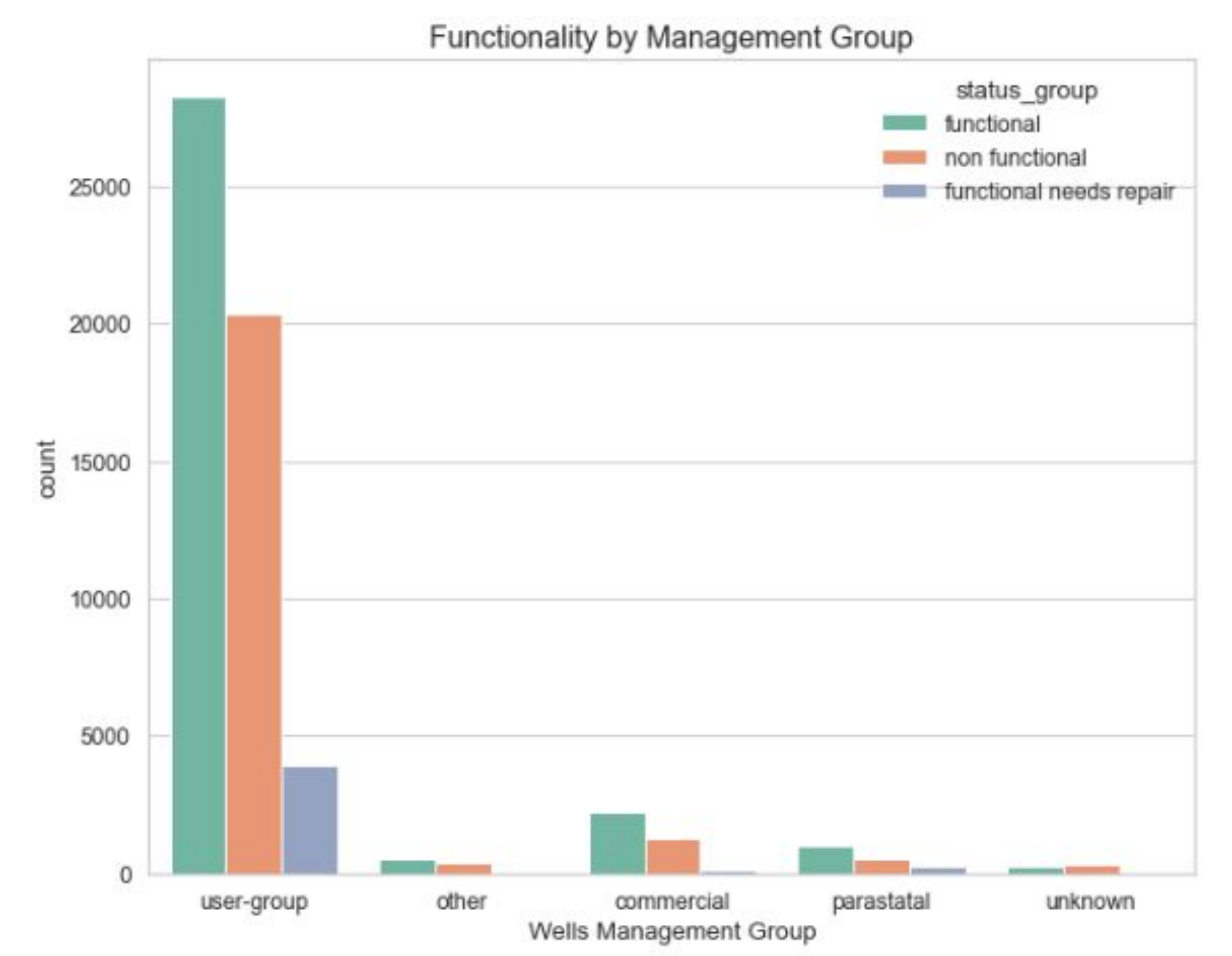


- Highlights a class imbalance problem which should be handled during modeling
- Aligns with prior research that 30–40% of rural water wells in Sub-Saharan Africa are non-functional at any given time.

Functionality by Extraction Type



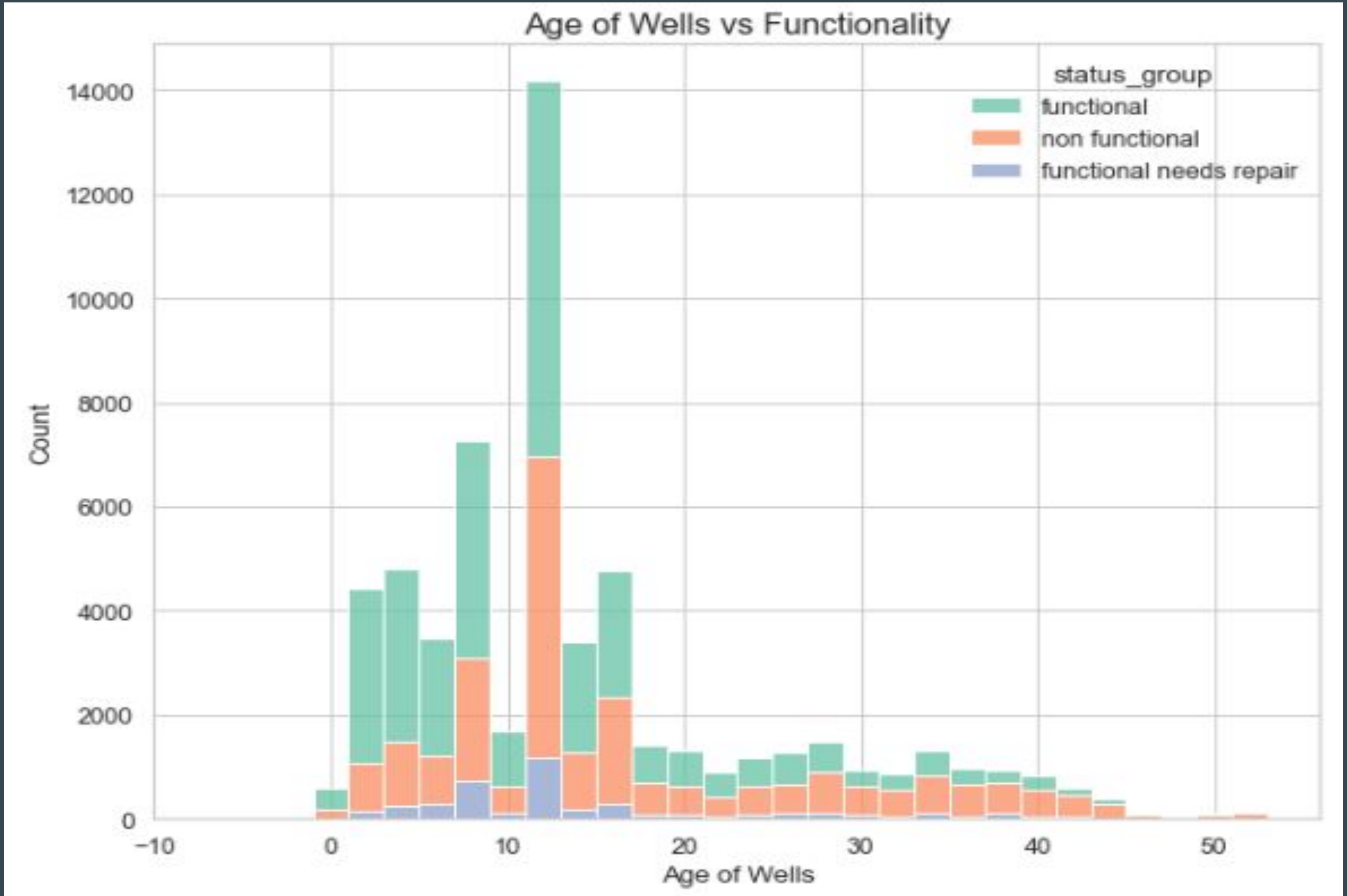
Technology choice matters. Motorpumps and “other” extraction types may have sustainability issues. NGOs and government should evaluate durability before funding new installations.



Community-based management is the norm but not always effective.

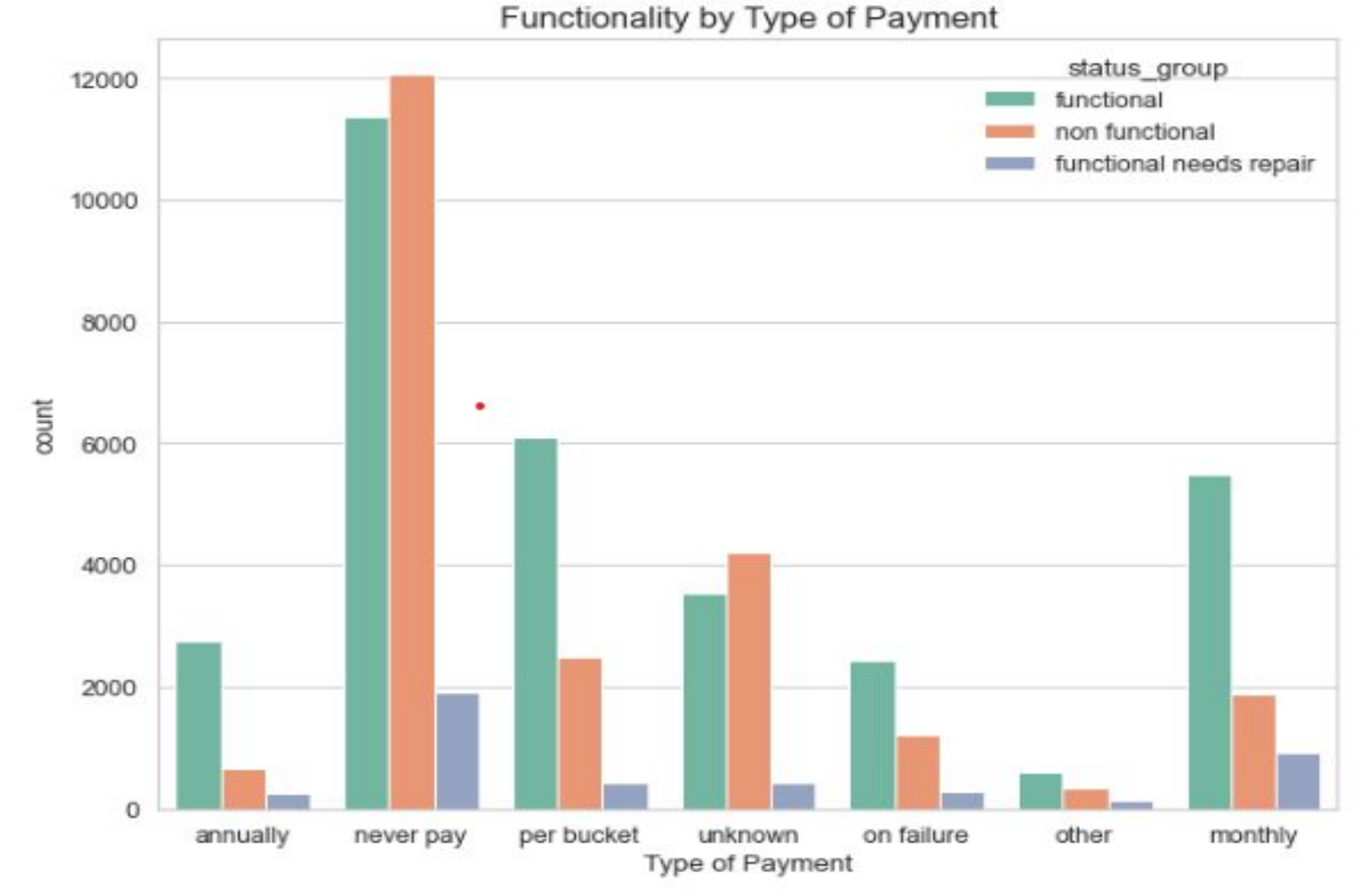
Highlights the need for capacity building and governance support for user-groups.

Age of Wells versus Functionality



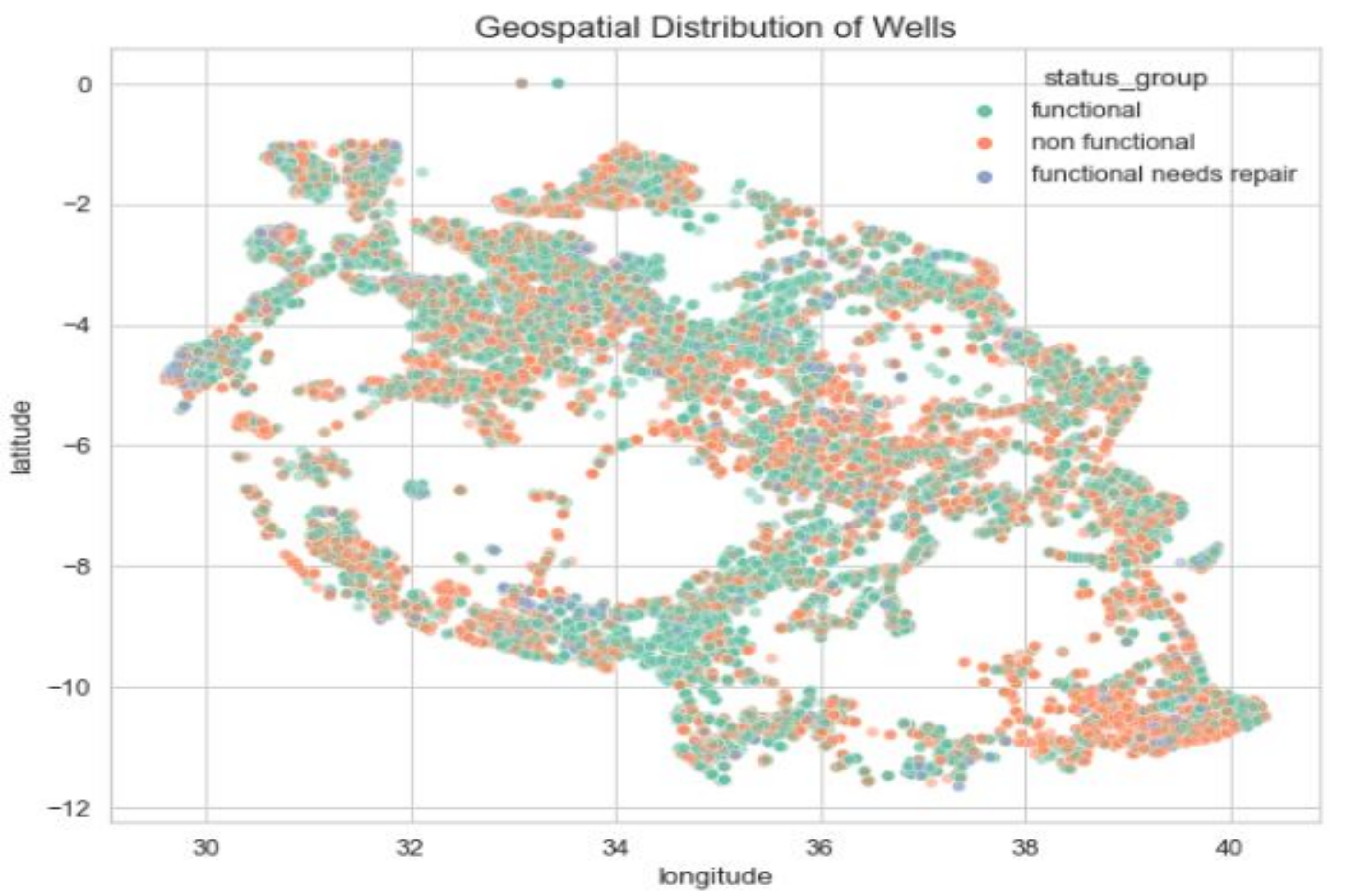
- Older wells degrade over time.
- Maintenance strategies should focus on wells aged 15+ years.

Type of Payment vs Functionality

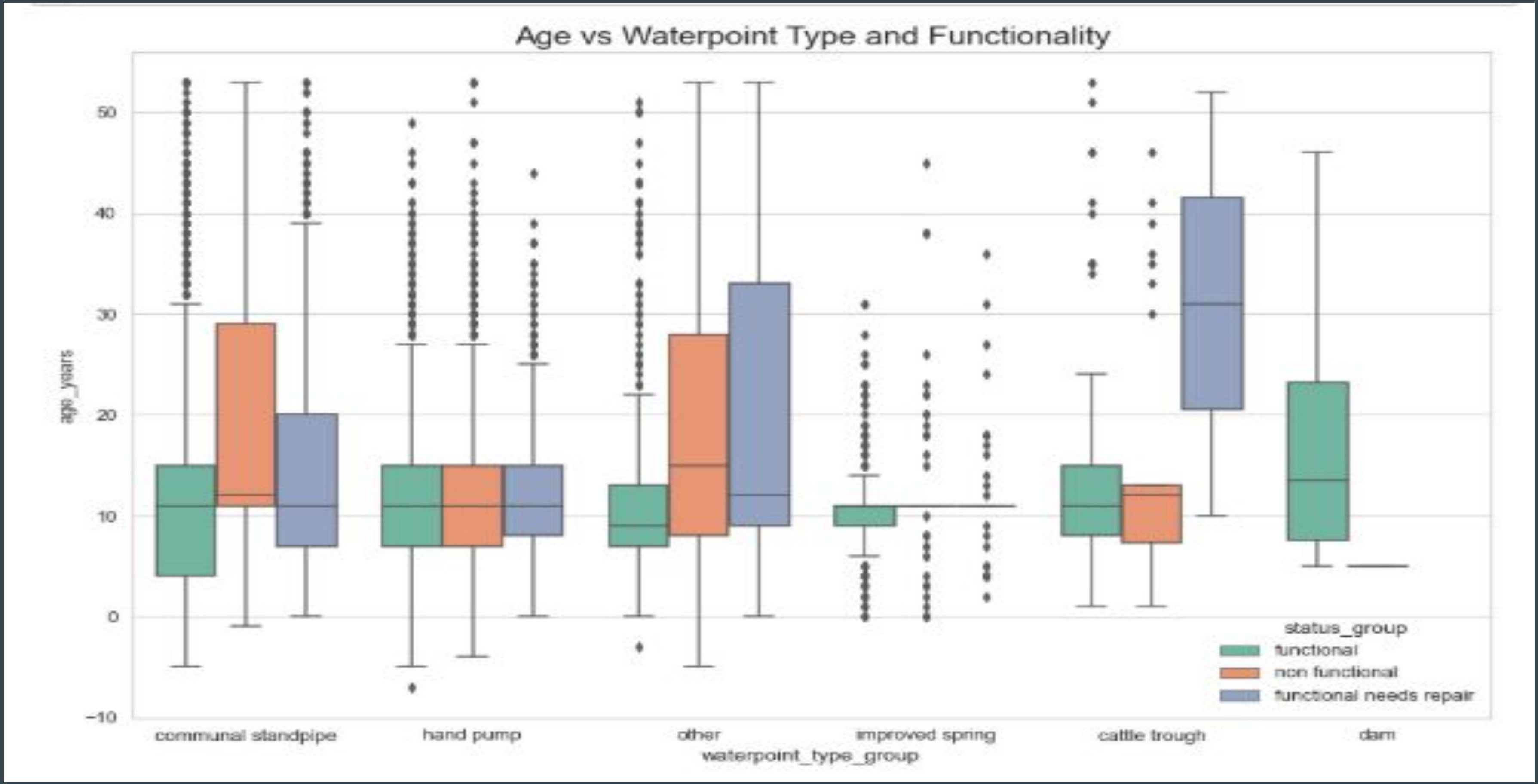


- Wells where communities “never pay” show high non-functionality.
- Monthly and per-bucket payments are associated with higher functionality.

Geospatial Distribution of wells



- Wells are spread across Tanzania, with clusters in central and northern regions.
- Geospatial clustering indicates regional vulnerabilities.



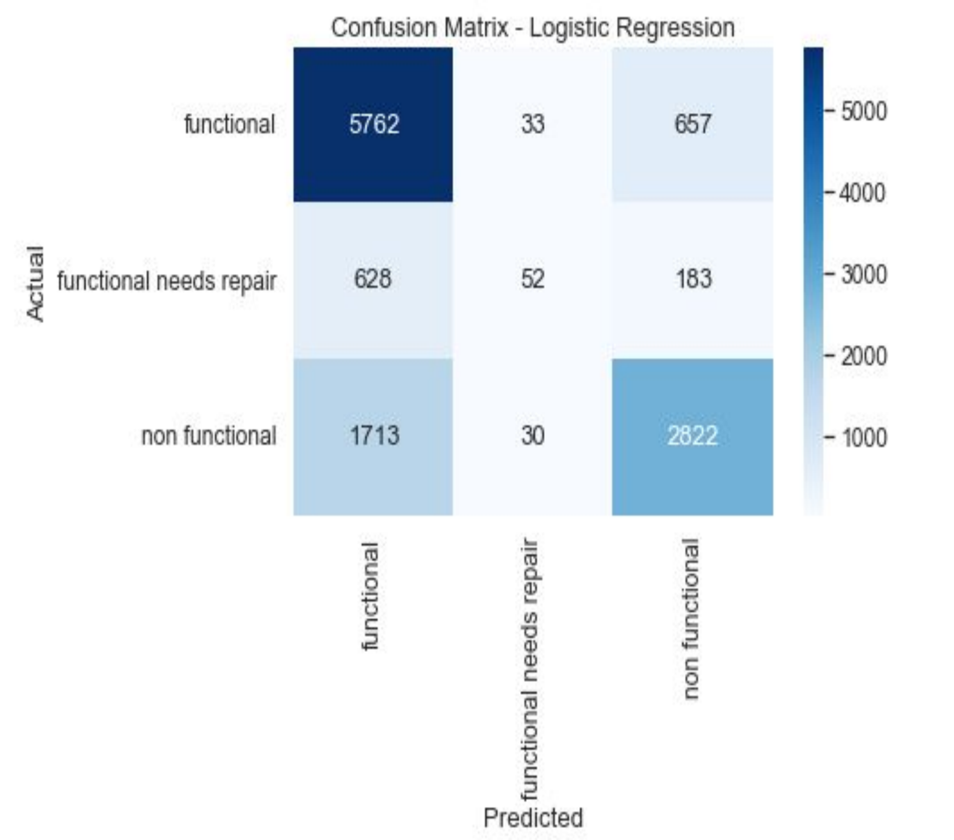
Technology choice and age jointly predict functionality.

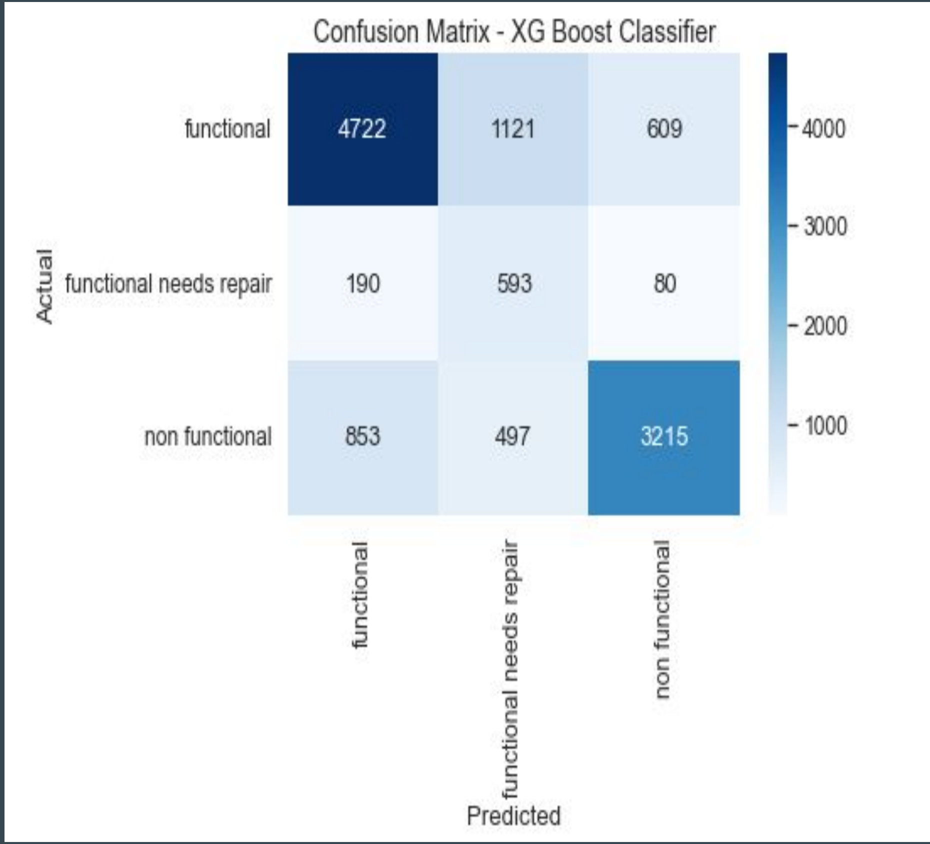
Modeling

Modeling Approach

- Tested multiple models: Logistic Regression, Random Forest, XGBoost, Gradient Boosting.
- Evaluated using accuracy, recall, and ability to detect failing wells.
- Focused on reducing false positives → avoiding labeling a broken well as functional.

Confusion Matrix - Logistic vs XGB





Results

Logistic Regression: Stable, but weak overall.

Random Forest: Strong accuracy, but risk of overfitting.

Gradient Boosting: Strong for functional wells, weak on failures.

XGBoost (Final Choice):

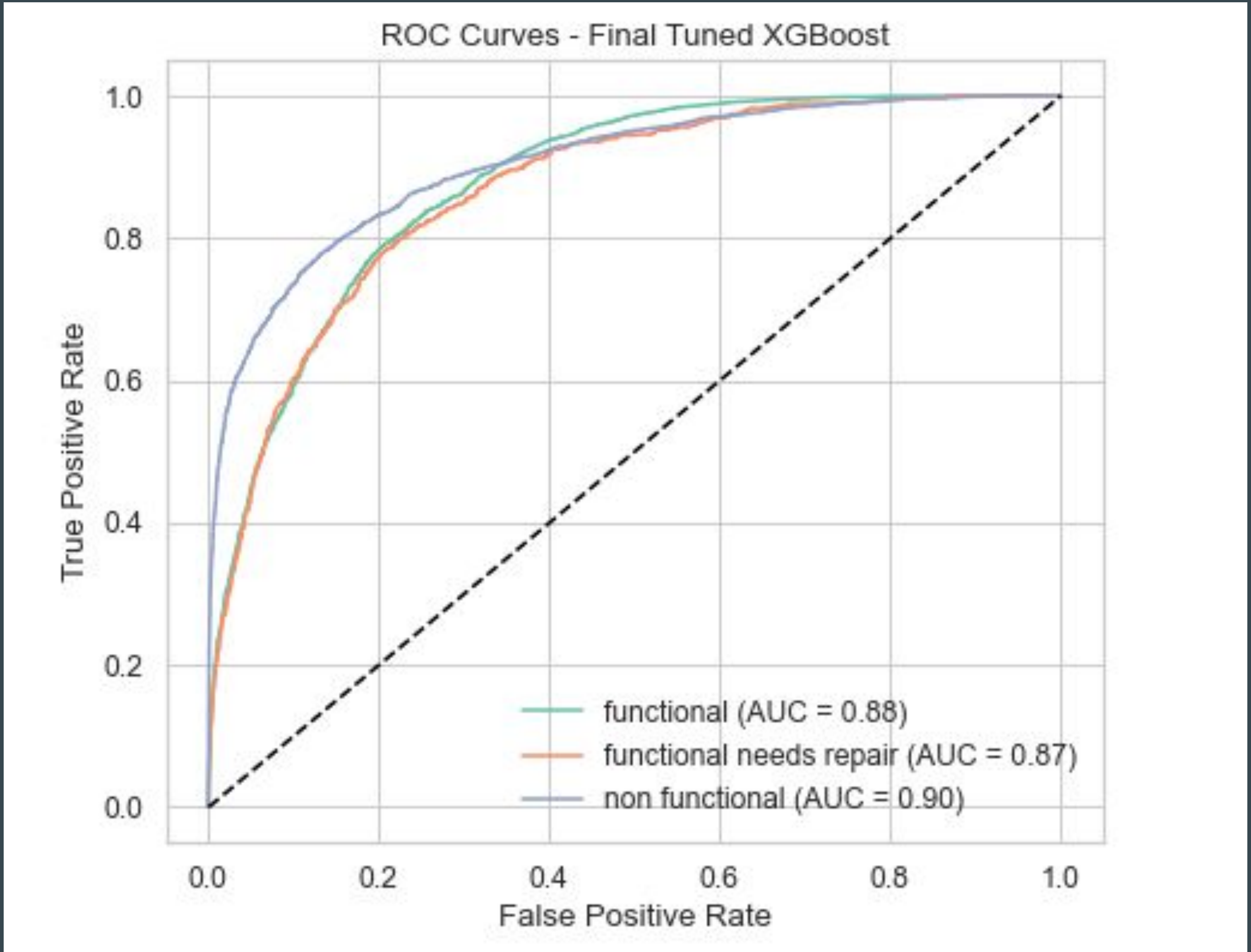
- Balanced performance
- Best at identifying wells “needing repair”
- Strong generalization to new data

Model Comparison: Logistic Regression vs Random Forest vs XGBoost vs Gradient Boosting

Metric	Logistic Regression	Random Forest	XGBoost	Gradient Boosting	Observation
Training Accuracy	72.98%	92.63%	76.17%	74.30%	RF is clearly overfitting (large gap). XGB and GBC show smaller gaps, indicating better generalization.
Validation Accuracy	72.69%	76.49%	71.80%	73.56%	RF achieves the highest validation accuracy, but XGB and GBC are more balanced.
Functional Class (Recall)	0.89	0.82	0.73	0.91	GBC best at identifying functional wells, RF slightly lower, LR strong, XGB weaker.
Functional Needs Repair (Recall)	0.06	0.44	0.69	0.09	XGB performs best on this minority class. RF is second-best. GBC and LR perform poorly.
Non Functional Class (Recall)	0.62	0.75	0.70	0.61	RF handles this class best. XGB performs well. GBC and LR slightly weaker.
ROC AUC (Functional)	0.82	0.87	0.86	0.83	RF and XGB stronger; GBC slightly weaker but solid.
ROC AUC (Needs Repair)	0.78	0.84	0.86	0.81	XGB best for minority class.
ROC AUC (Non Functional)	0.84	0.90	0.89	0.85	RF slightly ahead, XGB close, GBC behind.
Overfitting	Low (stable)	High (train–val gap)	Low–Moderate	Low–Moderate	RF overfits, while XGB and GBC maintain more stable performance.

Why XGBoost Wins

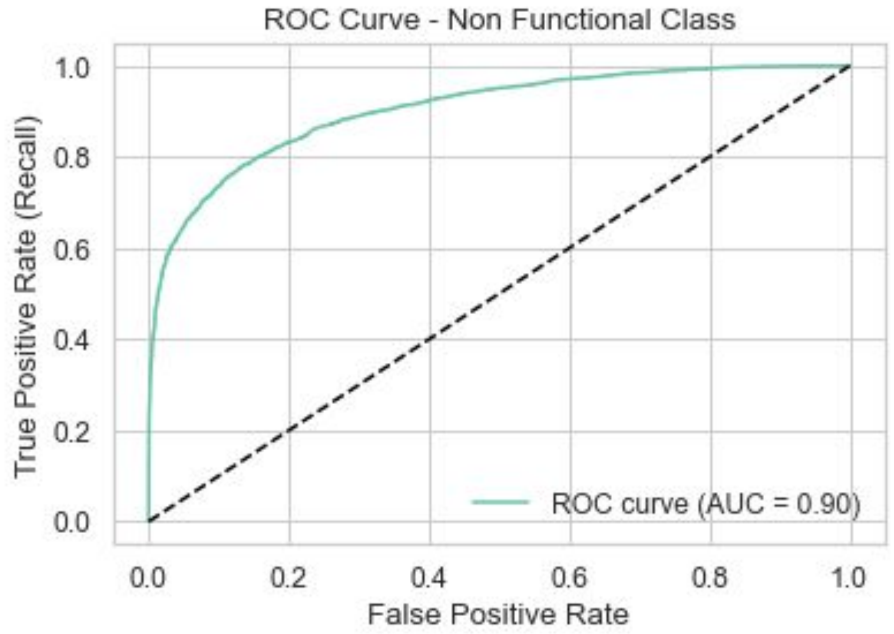
- Best balance between accuracy & recall for failing wells.
- Lower overfitting risk compared to Random Forest.
- Provides explainable feature importance (top drivers: pump type, management, water quantity, elevation).
- Aligns with the project’s mission: “Catch failing wells before communities go without water.”

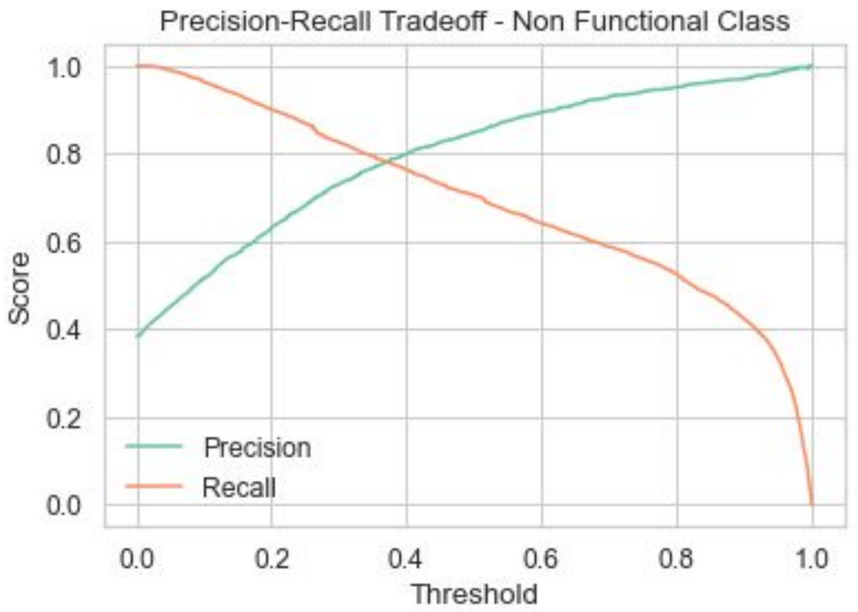

ROC Curves - Final Tuned XGBoost
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
functional (AUC = 0.88)
functional needs repair (AUC = 0.87)
non functional (AUC = 0.90)

Model Comparison: Baseline XGBoost vs Tuned XGBoost

Metric	Baseline XGBoost	Tuned XGBoost	Observation
Training Accuracy	76.17%	76.17%	Both similar, tuning didn't overfit further.
Validation Accuracy	71.80%	74.00%	Tuning improved validation accuracy by ~2.2%.
Functional Class (F1-score)	0.77	0.79	Small but consistent improvement.
Functional Needs Repair (F1-score)	0.39	0.41	Recall improved (0.69 → 0.64 slightly down), but F1 remained stable.
Non Functional Class (F1-score)	0.76	0.78	Slightly stronger predictive performance after tuning.
Macro Avg F1	0.64	0.66	Improvement shows more balanced classification across classes.
ROC-AUC (macro avg)	0.870	0.883	Clear improvement in discriminative ability with tuning.
Confusion Matrix	More false positives (non-functionals predicted as functional)	Slightly reduced false positives, higher true positives for non-functional	Indicates improved balance after tuning.

Threshold Tuning





Comparison: Before vs After Threshold Tuning

Metric	Before Threshold Tuning	After Threshold Tuning (0.35)	Observation
Accuracy	76%	74%	Small drop in accuracy (expected when optimizing for recall).
Functional Precision/Recall	P: 0.82 / R: 0.79	P: 0.85 / R: 0.71	Precision improved, recall slightly decreased.
Needs Repair Precision/Recall	P: 0.33 / R: 0.56	P: 0.31 / R: 0.62	Recall improved (captures more at-risk wells), precision slightly lower.
Non-Functional Precision/Recall	P: 0.81 / R: 0.76	P: 0.77 / R: 0.79	Better balance achieved: fewer false positives of broken wells predicted as functional.
ROC-AUC (macro)	0.879	0.883	Small improvement in discriminative power.

Conclusion

- Most wells are functional, but ~45% fail or need repair.
- Failures cluster in certain regions (Iringa, Shinyanga, Morogoro).
- Pump type, management, elevation, and population are key risk factors.
- Predictive models achieved >70% accuracy, confirming failures are predictable.
- Random Forest & XGBoost performed best; both provide valuable insights.

Recommendations

- Preventive Maintenance
- Target High-Risk Regions
- Strengthen Local Management
- Smarter Resource Allocation
- Policy Integration

Limitations

- Data Quality Issues
- Temporal Gap - Does not capture seasonality and recent water availability changes
- Class Imbalance
- Model Generalisation - Future changes in infrastructure or climate may affect accuracy

Next Steps

- Integrate weather, groundwater and socio-economic data.
- Deploy a dashboard - To show predictions
- Cost-sensitive learning - Penalize misclassifying failing wells.


Any
Question

THANK YOU

•••

Data Scientist: Huldah Chepkoech
LinkedIn Profile: Huldah (Chepkoech) Rotich