Predicting Tanzania's Waterpoint Reliability: A Data-Driven Approach to Operational Status

By Roy Njuguna
March 7, 2025

SUMMARY

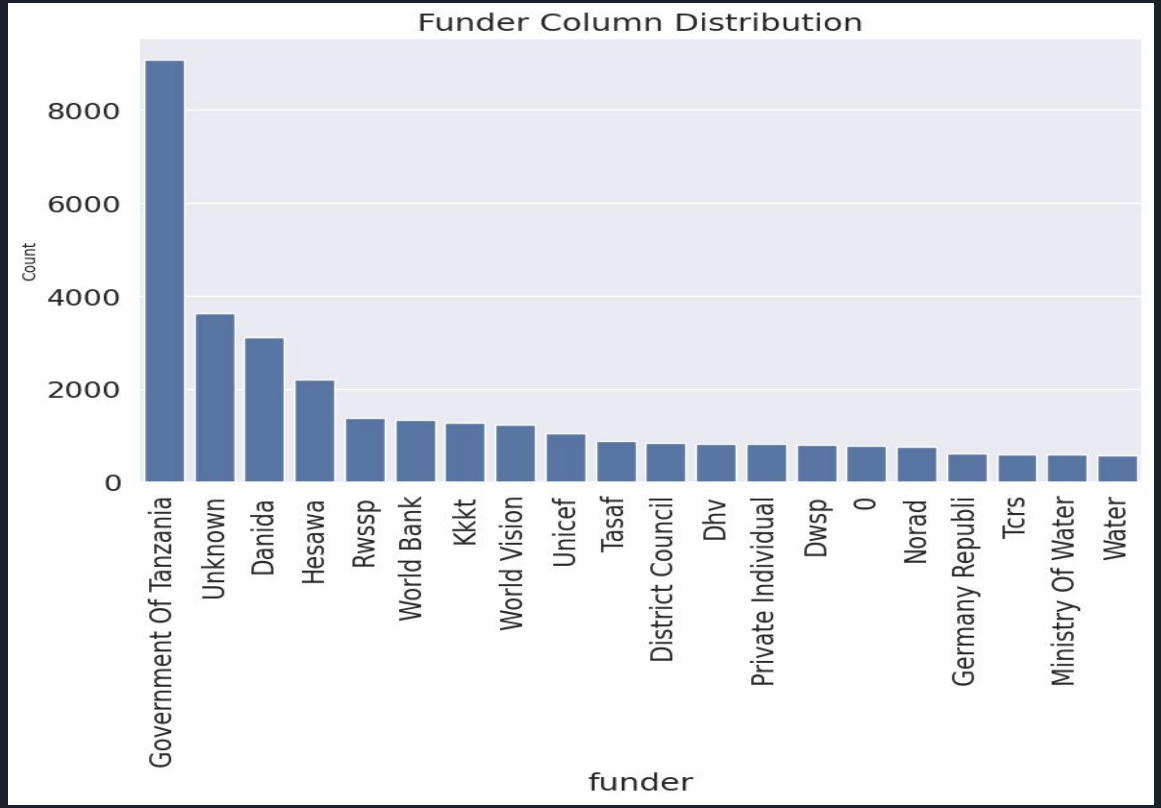
This project focuses on addressing water access challenges in Tanzania, where many rural and remote communities struggle to maintain functional water points. By leveraging data and machine learning, the goal is to predict the operational status of these water points, supporting efforts to improve access to safe and reliable water sources.

BUSINESS UNDERSTANDING

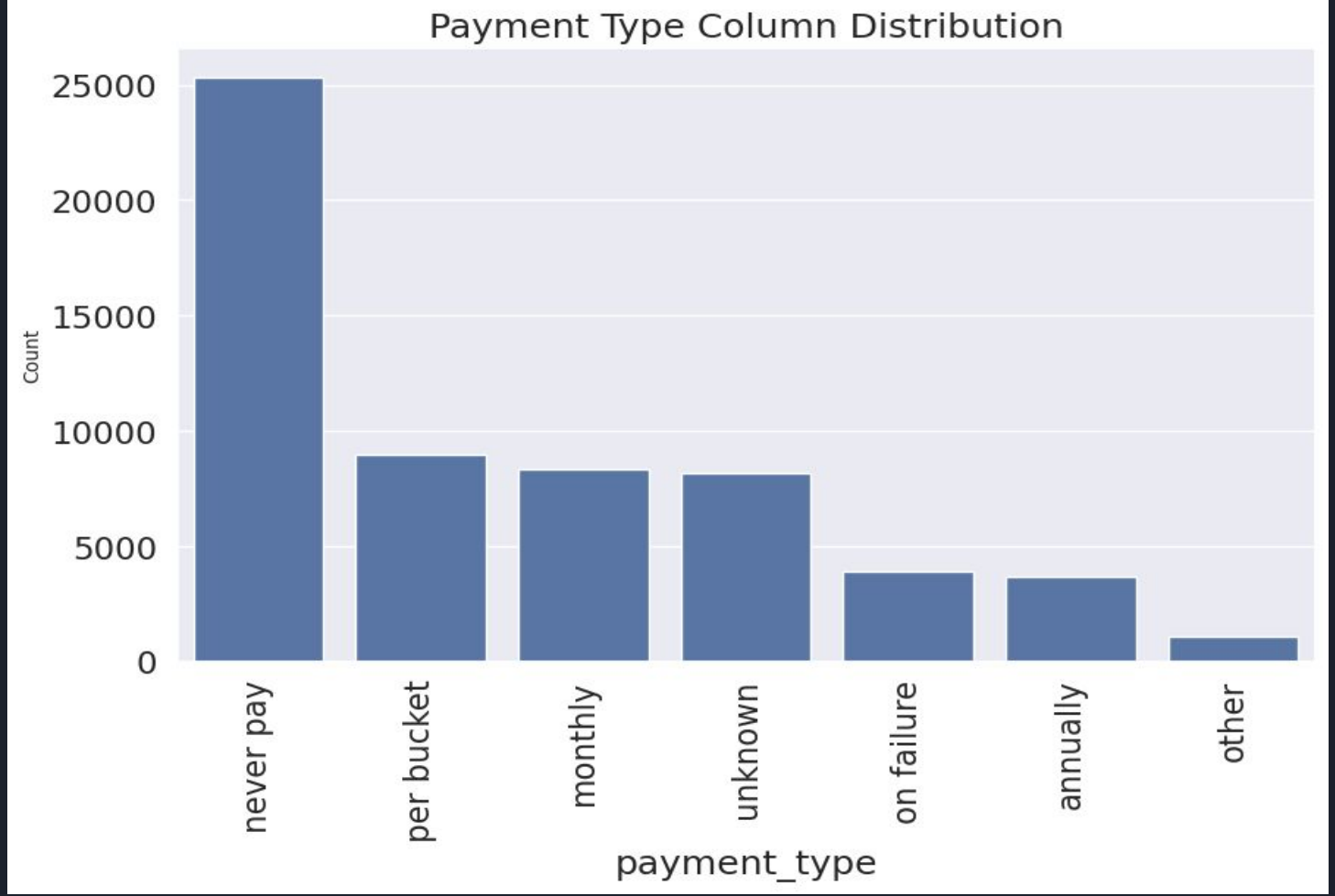
Reliable water access is vital for community well-being in Tanzania, but many waterpoints face operational issues. This project develops a predictive model to assess their status, enabling proactive maintenance and improved resource management. By identifying failing pumps early, we enhance access to clean water and support sustainable water management.

DATA UNDERSTANDING

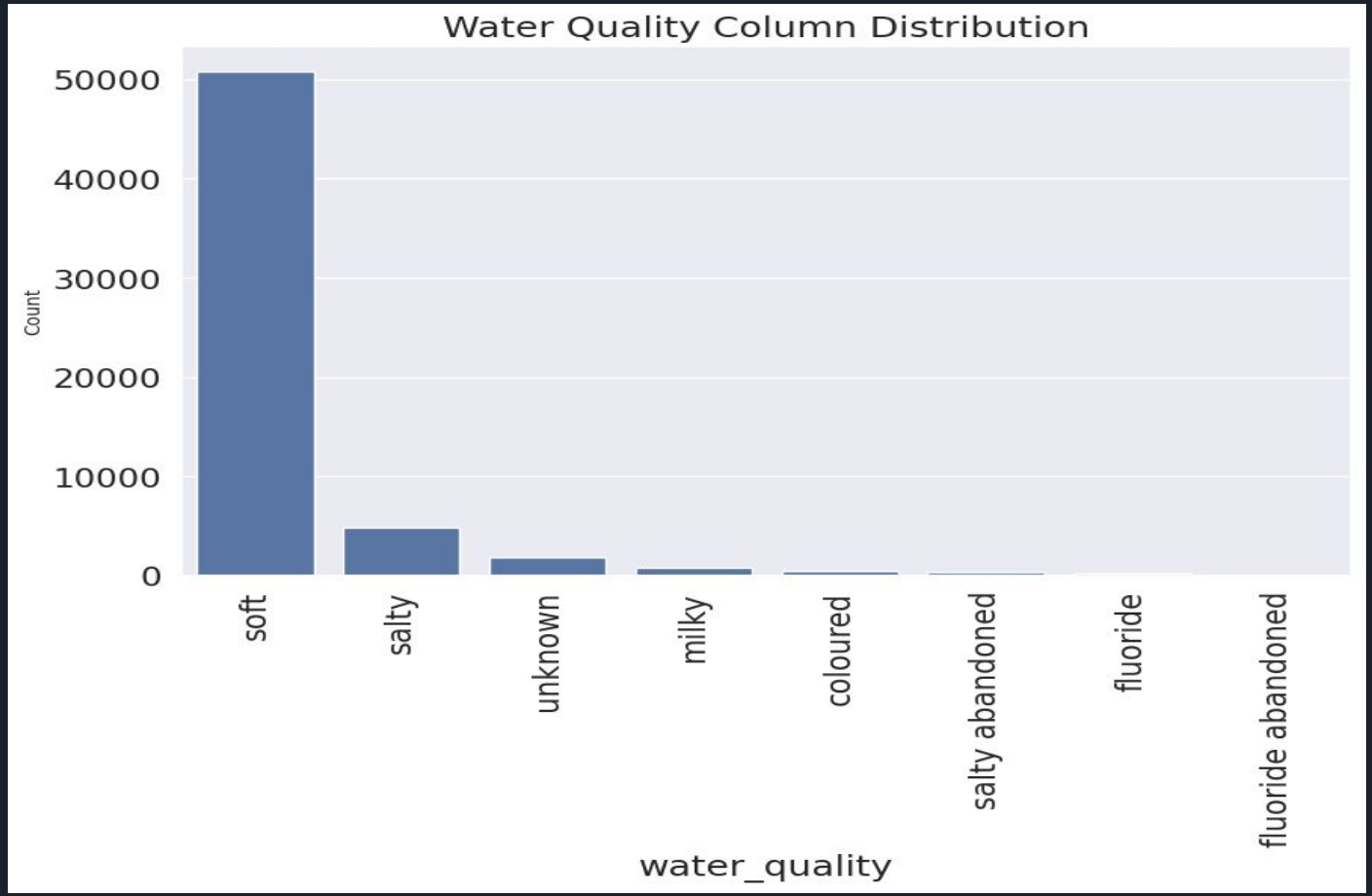
The dataset contains detailed information on waterpoints across Tanzania, capturing a variety of features that may influence the operational status of each waterpoint. These features range from geographic details to waterpoint characteristics, such as funding source, installation details, water quality, and management practices.



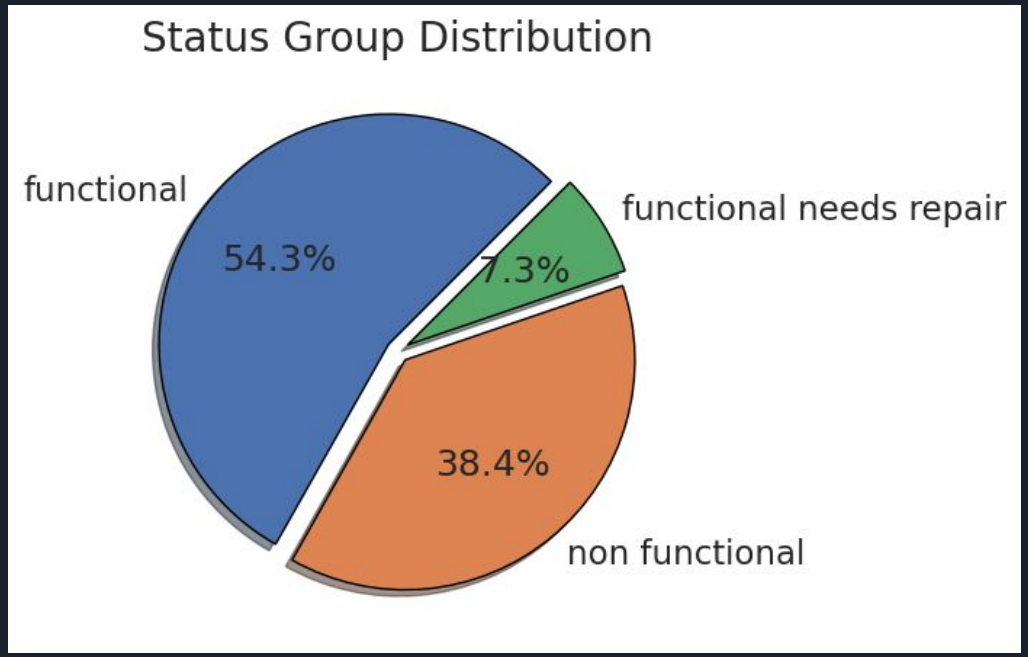
Most wells in Tanzania are funded by the Government of Tanzania



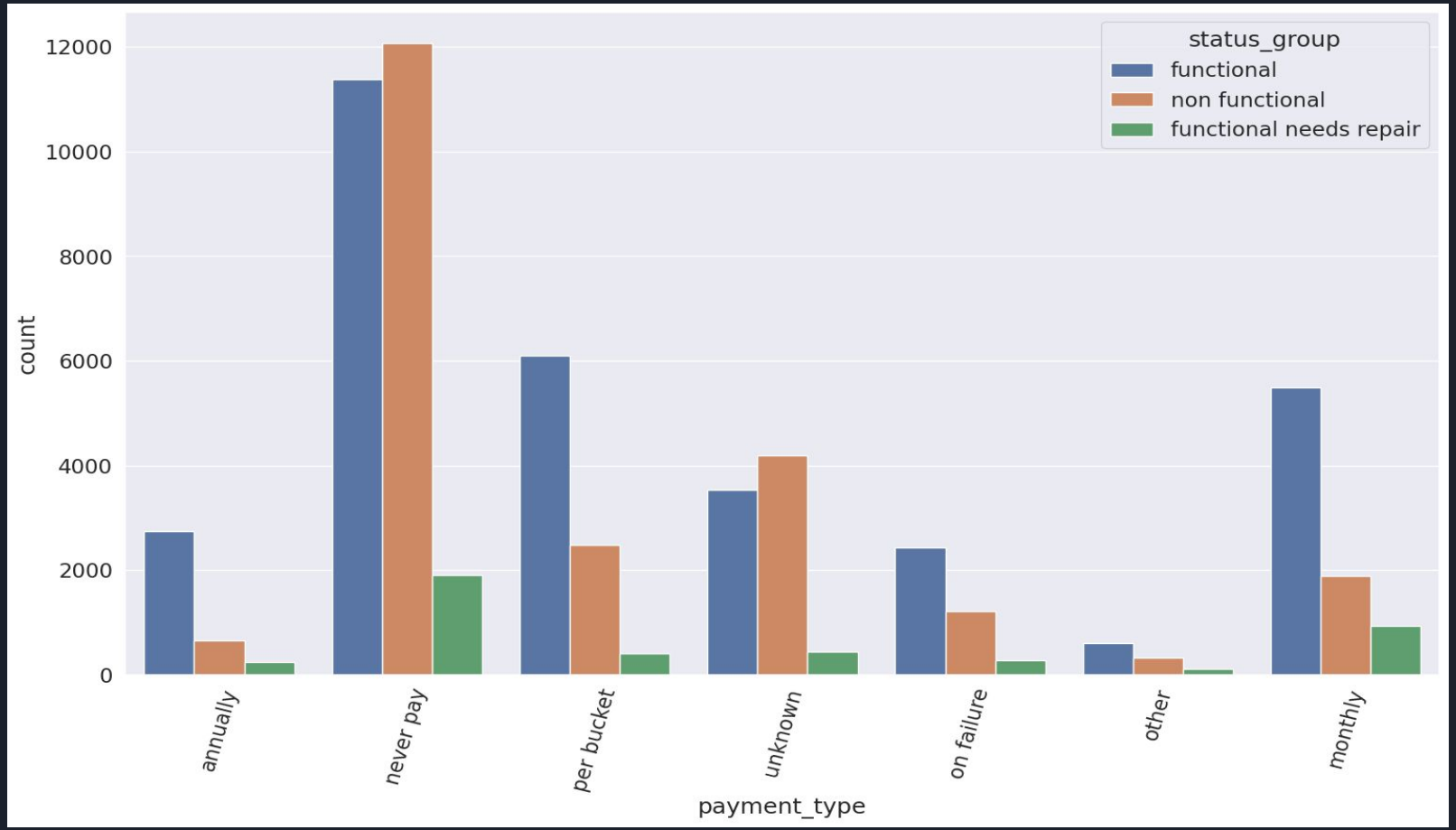
Most people access water from wells without payment, likely because these wells are intended for community benefit rather than profit.



Soft water lacks calcium and magnesium salts, which in excess can be harmful to both health and homes. Examples include rainwater and distilled water. Our visualization shows that communities in Tanzania primarily consume soft water.



Our pie chart indicates that around 55% of water pumps are fully functional, 7% are operational but require repairs, and 38% are non-functional.



Most of the functional and non functional water pumps are never paid for, again this might be because of the fact that they are communal

MODELLING

We use XGBoost to predict water pump status in Tanzania due to its efficiency and superior performance over Random Forest. With **79.91% accuracy** (**7% higher than RF**), better recall (**76% vs. 54%**), and stronger generalization (**Macro F1: 0.68, Weighted: 0.79**), XGBoost ensures balanced predictions.

To optimize performance, we fine-tune **n_estimators, max_depth, learning rate, alpha, and lambda** using **Grid or Random Search with cross-validation**, preventing overfitting and improving accuracy.

EVALUATION

We assess the model using **accuracy (79.91%)**, **precision**, **recall (76% for Class 1 vs. 54% in RF)**, and **F1-score (Macro: 0.68, Weighted: 0.79)**. The **confusion matrix** helps analyze misclassifications, ensuring balanced predictions. Comparing these metrics across datasets evaluates generalization and areas for improvement.

RECOMMENDATIONS

Based on the model's performance and evaluation metrics, the following recommendations can enhance water point operational status predictions:

1. **Feature Engineering:** Incorporate additional relevant features, such as weather conditions, population density, and infrastructure data, to improve model accuracy.
2. **Data Quality Improvement:** Address missing or inconsistent data through imputation techniques and better data collection methods to enhance model reliability.
3. **Model Optimization:** Experiment with other ensemble methods, such as LightGBM or CatBoost, to compare performance and potentially improve results.
4. **Regular Model Updates:** Retrain the model periodically with updated data to adapt to changing conditions and maintain accuracy.
5. **Deployment and Monitoring:** Implement the model in a real-world system with continuous monitoring to assess performance and make necessary adjustments based on new data.

These steps will help refine the model, ensure better predictions, and support efforts to improve water access in Tanzania.

NEXT STEPS

To further improve the project and ensure its practical impact, the following next steps are recommended:

1. **Expand Data Sources:** Integrate additional datasets, such as satellite imagery and real-time sensor data, to enhance model insights.
2. **Deploy the Model:** Implement the trained model in a user-friendly application or dashboard for decision-makers to monitor water point functionality.
3. **Conduct Field Validation:** Collaborate with local authorities and organizations to validate model predictions with real-world observations.
4. **Automate Data Pipeline:** Establish a continuous data collection and processing pipeline to keep the model updated with the latest information.
5. **Scale the Approach:** Explore the feasibility of applying the model to other regions facing similar water access challenges.

CONCLUSION

This project successfully predicts the operational status of water pumps in Tanzania using **XGBoost**, which outperformed **Random Forest** based on **accuracy (79.91%)**, **precision**, **recall (76% vs. 54%)**, and **F1-score (Macro: 0.68, Weighted: 0.79)**. By leveraging these insights, proactive maintenance can be improved, ensuring better water access and sustainability for communities.

Thank you for reviewing this project. Your time, insights, and support in improving data-driven solutions for clean water access are greatly appreciated.

Email: roynjuguna222@gmail.com