The Tanzania Water Classification Model

A Machine Learning Classification Project

By Ryan Posternak

5 sections in 5 minutes

Agenda

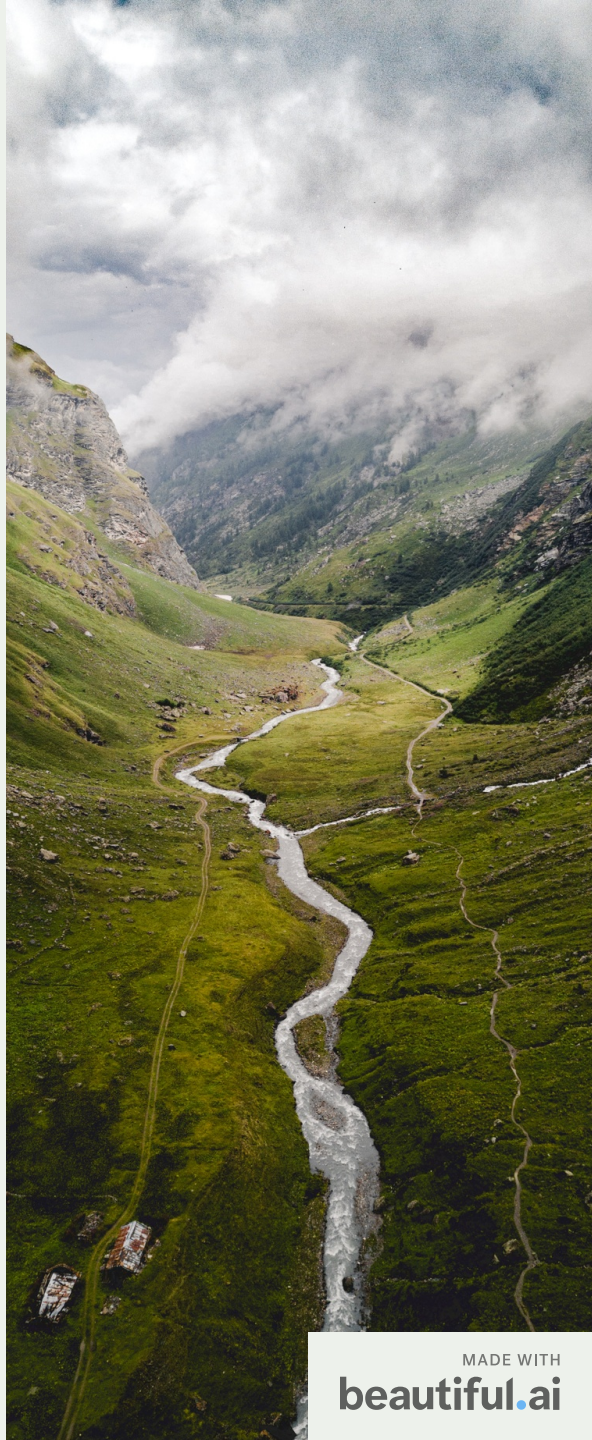
1 | GOALS & BUSINESS UNDERSTANDING

2 | THE DATA

3 | THE MODELS

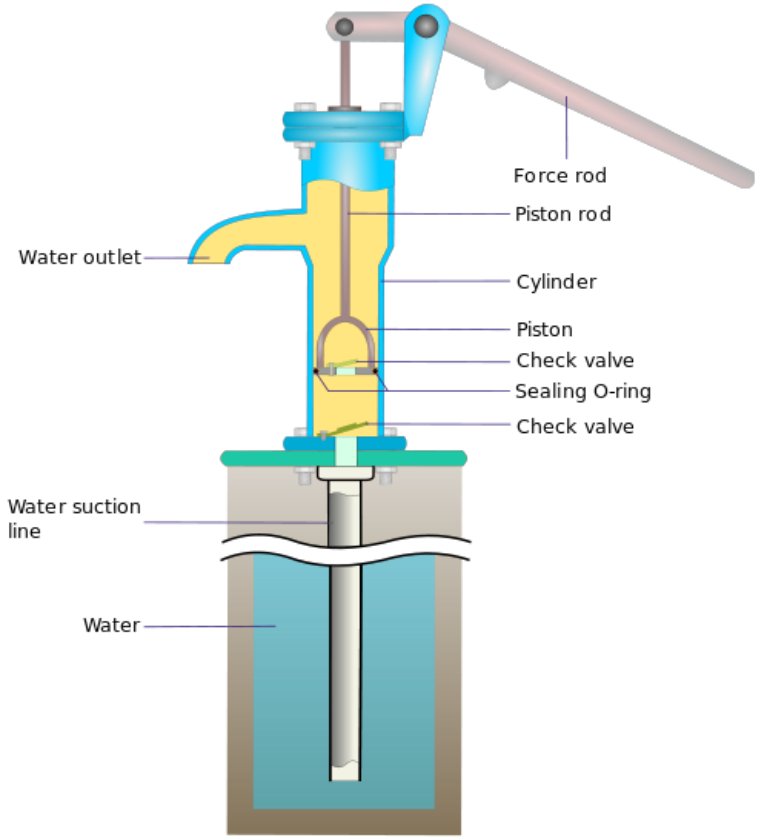
4 | EVALUATION & INFERENCES

5 | CONCLUSIONS & RECOMMENDATIONS



What are the goals?

Business Understanding



WATER CRISIS

As much as half the country — over 20 million people — do not have reliable access to safe drinking water

PREDICTIVE VALUE: BEING PROACTIVE, NOT REACTIVE

By knowing in advance the proportion of pumps that will breakdown, TMW can plan ahead and save money

WATER IS EXPENSIVE

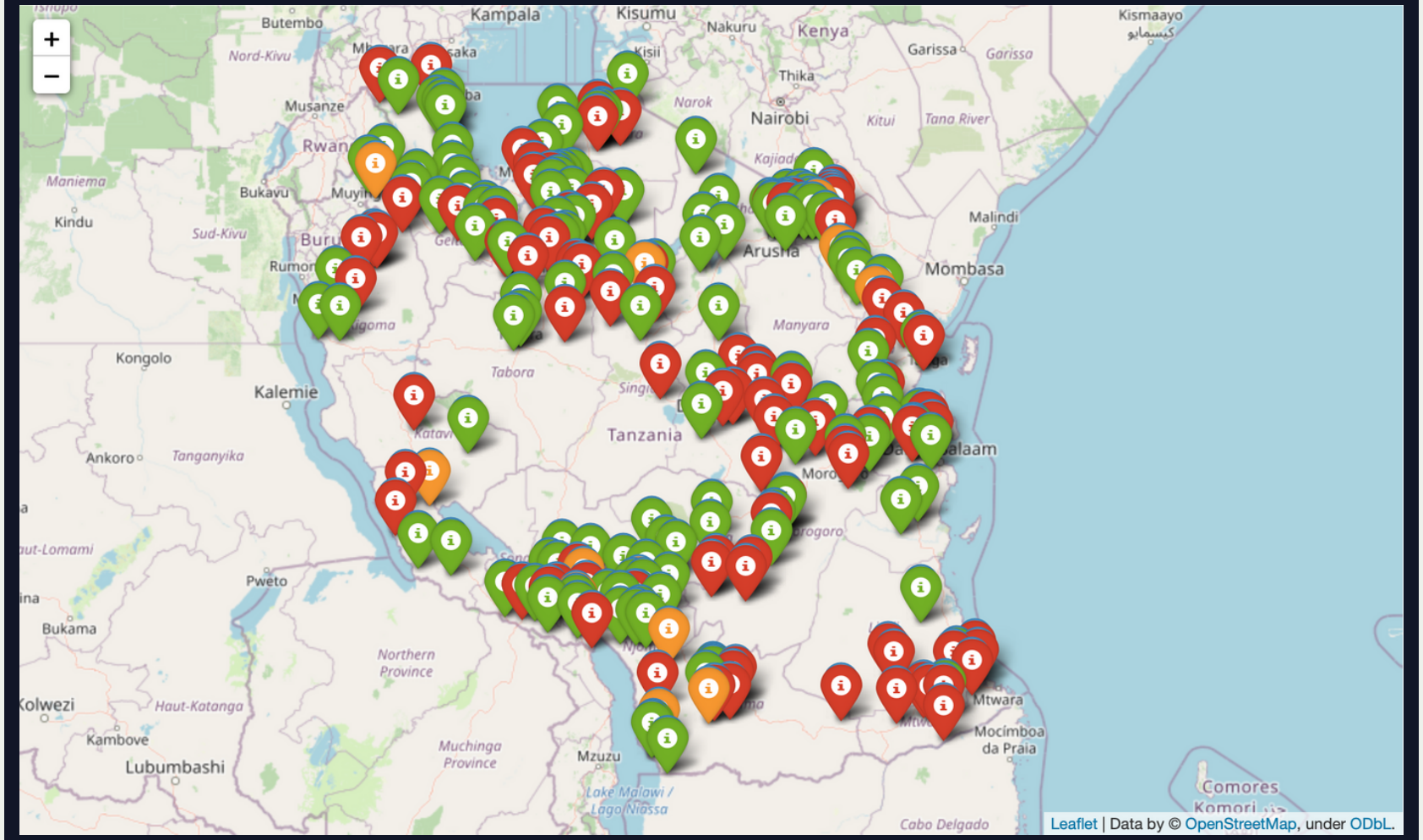
TMW's '22 budget: $294M USD

~2% of entire budget of government of Tanzania

INFERENTIAL VALUE: REDUCE SOURCES OF PUMP BREAKDOWN

By analyzing the most important features, inferences can be made on what factors tend to lead to water pump failure

Data Understanding



- **AGGREGATED DATA FROM THE TANZANIA MINISTRY OF WATER**

- **TAARIFA WATERPOINTS DASHBOARD**

  Taarifa: An open source platform for the crowd sourced reporting and triaging of infrastructure related issues

- **59,400 ROWS, 40 COLUMNS**

  3 class labels:

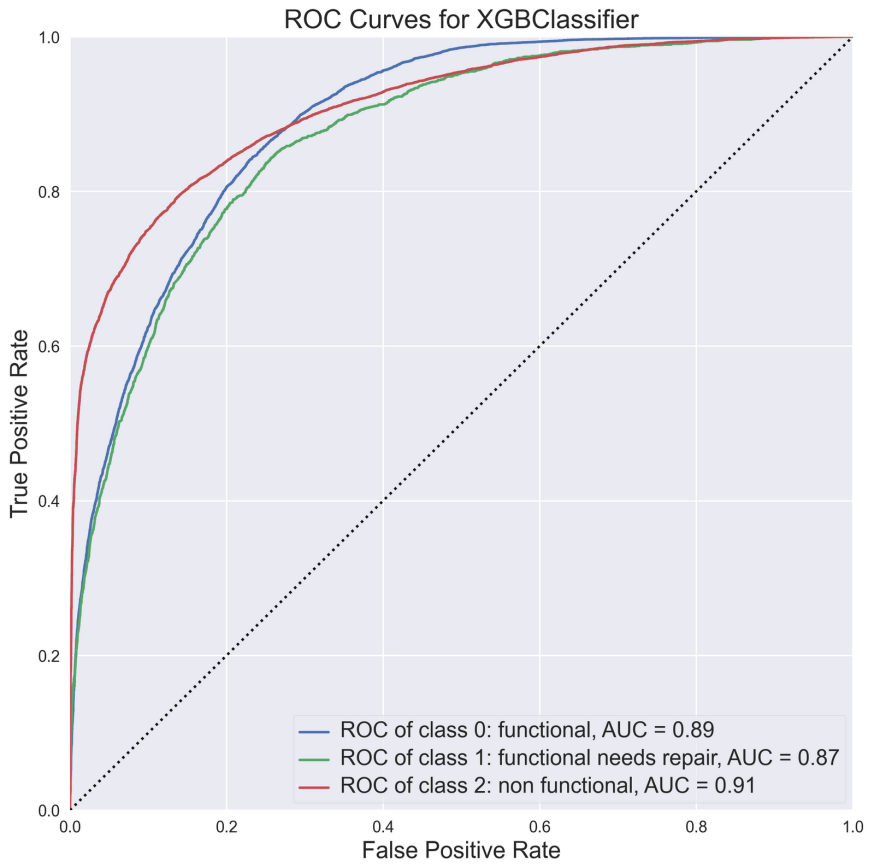
“functional”:	54.3%
“functional needs repairs”:	7.3%
“non functional”:	38.4%

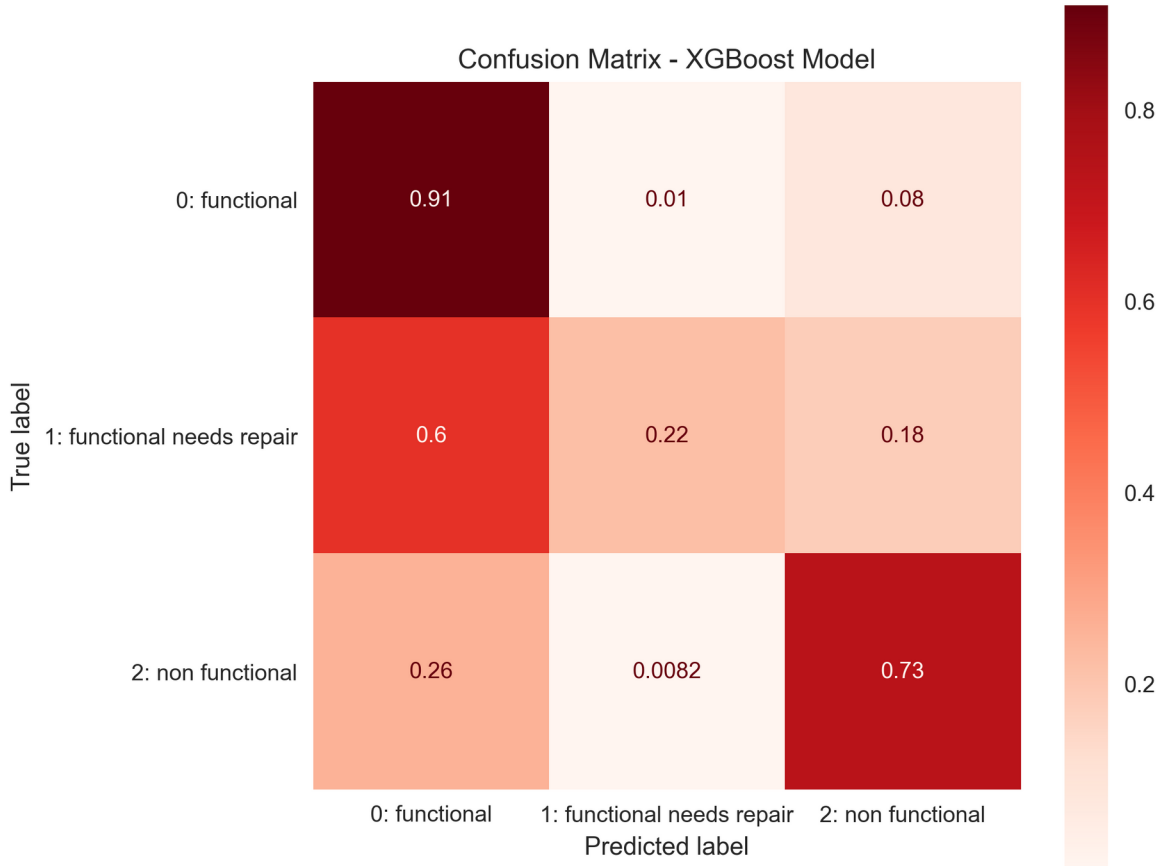
Models, Ranked by Test Accuracy Score

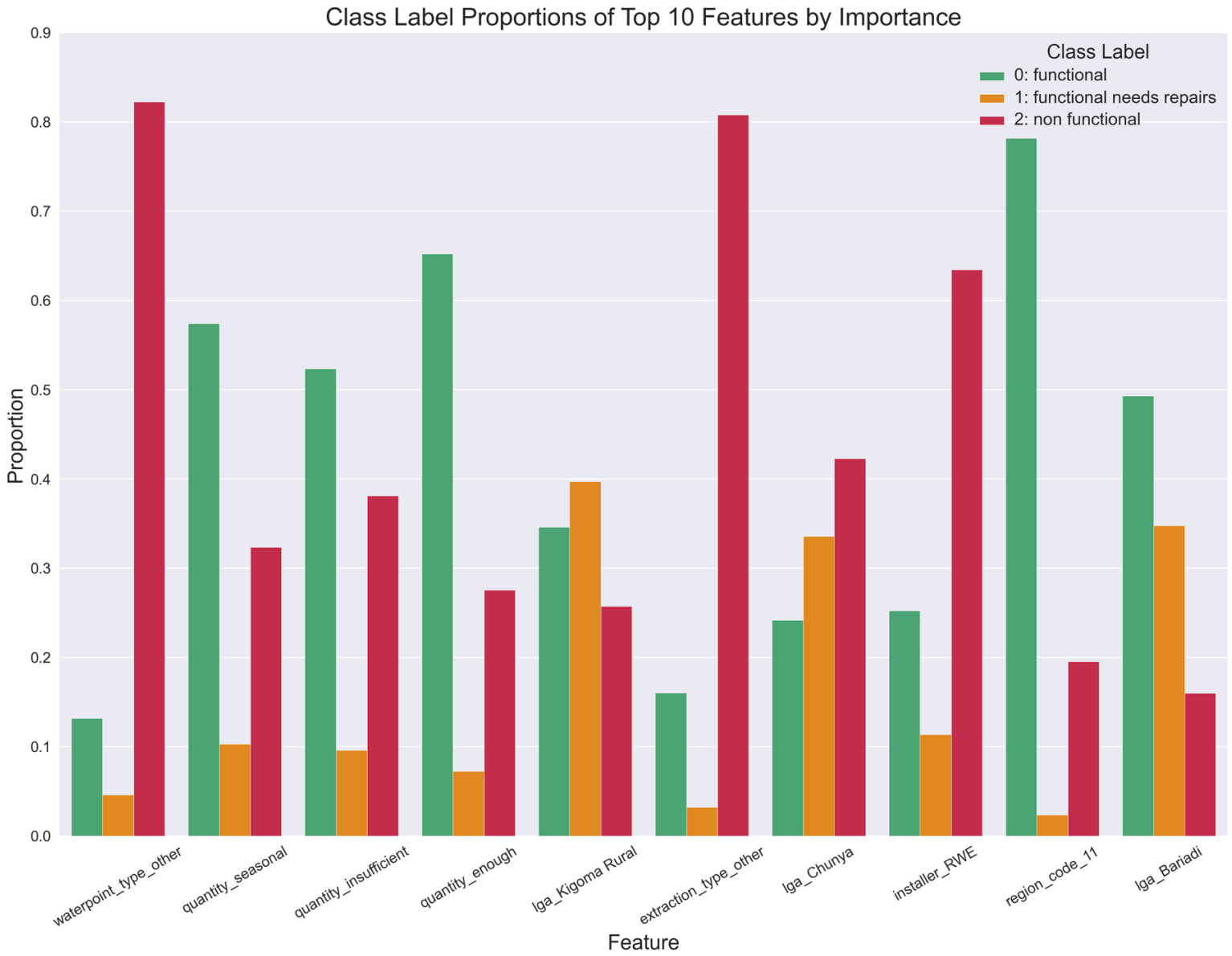
Model	Training Accuracy	Cross-Validation Accuracy	Test Accuracy
Model 7.1: XGBoost	**84.4%**	**78.5%**	**79.2%**
Model 5.1: Gradient Boosting	93.9%	77.9%	78.7%
Model 6.1: Random Forest	84.2%	77.8%	78.6%
Model 4.1: AdaBoost	95.5%	76.2%	76.8%
Model 2.1: KNN	81.2%	74.6%	76.4%
Model 3.2: Decision Tree 2	83.5%	75.1%	75.9%
Model 1.2: Logistic Regression 2	75.9%	75.2%	75.5%
Model 1.1: Logistic Regression 1	75.8%	75.2%	75.5%
Model 3.1: Decision Tree 1	95.6%	73.7%	74.7%
Model 0: Baseline Dummy	54.3%	54.3%	54.3%

The Metrics That Matter

Evaluation: XGBoost Classifier Model Metrics







Inferences

waterpoint_type: other & **extraction type: other** associated with high non-functionality rates

installer: RWE associated with high non-functionality rate

region_code: 11 is doing something right

Thank you

Email: rposternak@yahoo.com

GitHub: github.com/rjpost20

LinkedIn: linkedin.com/in/ryanposternak