Tanzania Water Wells Classification Project

Morgan Nash
August 2025

Business Understanding:

Problem:

- Tanzania experiencing water crisis
- Widespread non-functional water pumps
- Disruptions in clean water access

Objective:

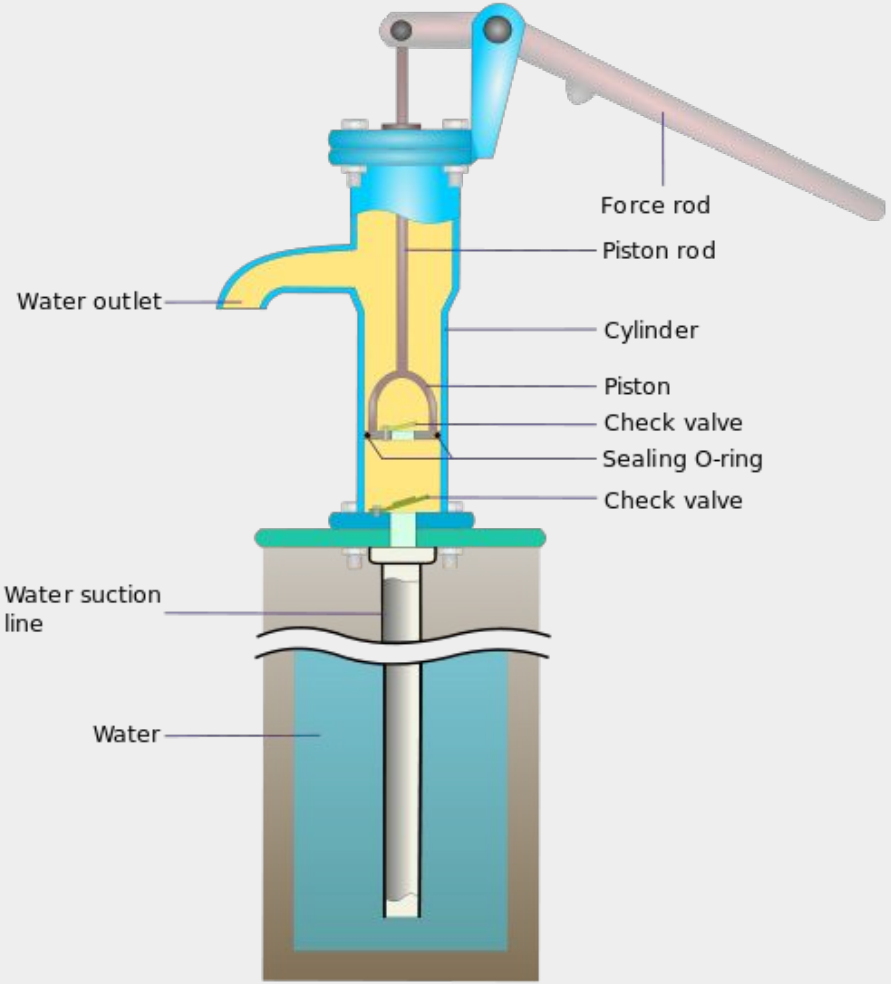
- Help stakeholders improve maintenance efficiency
- Move from a reactive to a proactive maintenance strategy

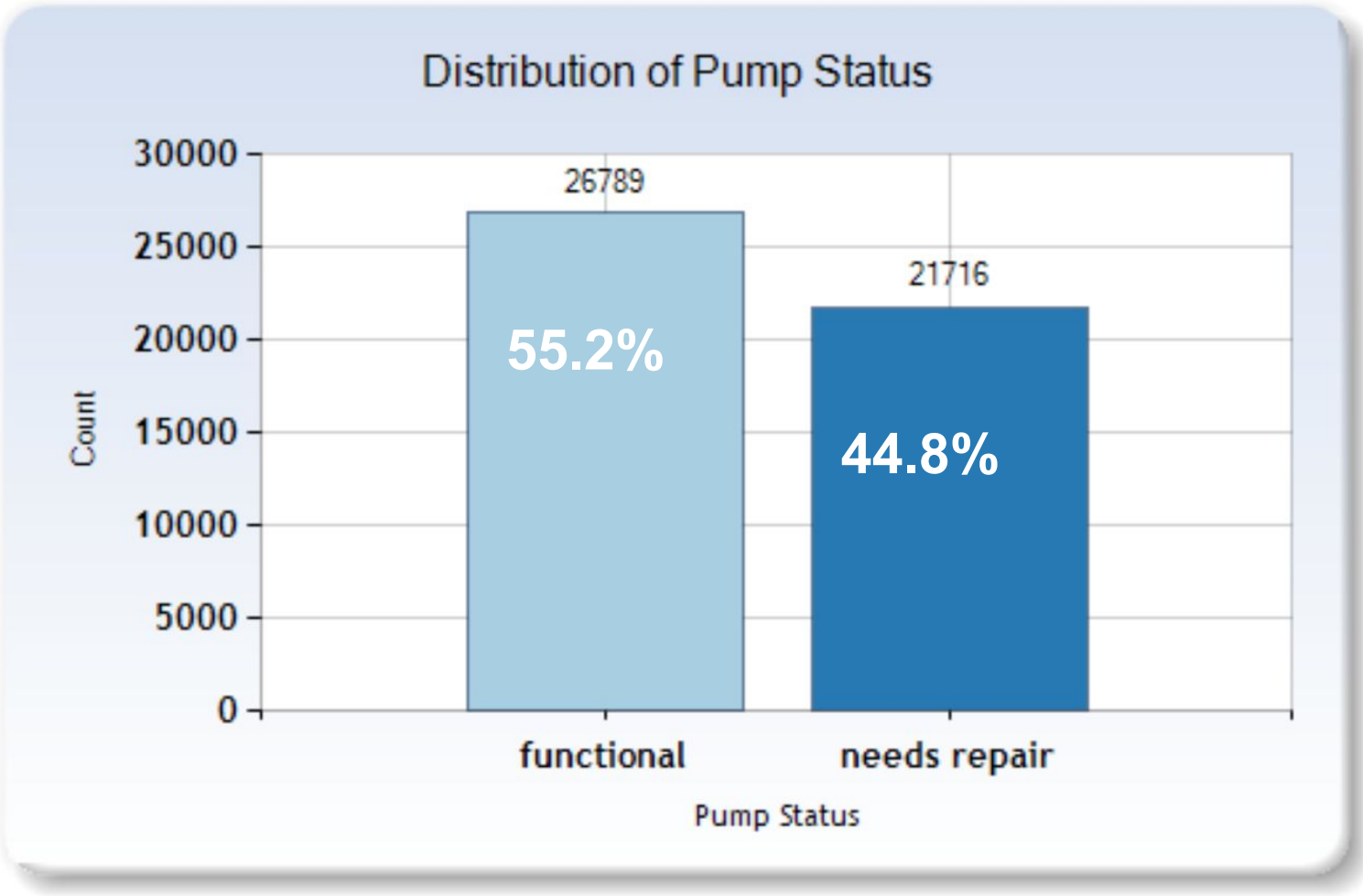
Stakeholders:

- Non-Government Organizations (NGOs) focused clean water access
- Tanzanian Government

Data Understanding:

- Data from the Tanzania Ministry of Water, accessed via DrivenData.org
- 59,400 water well records
- 41 features, including location, construction date, water source, and quantity of water available to the pump
- Good choice because it's extensive and directly relevant




Distribution of Pump Status
30000
25000
20000
15000
10000
5000
0
Count
26789
21716
55.2%
44.8%
functional
needs repair
Pump Status

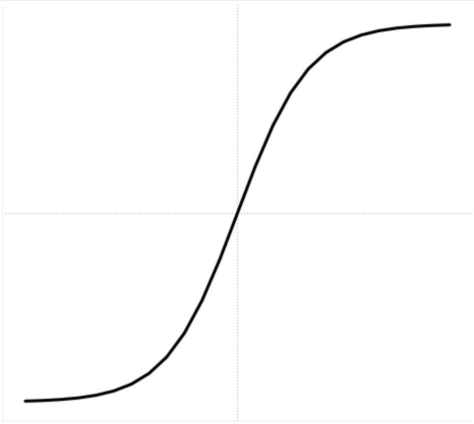
Why Recall?

- Cost of a mistake is not equal
- False Positive: model predicts a pump needs repair when it doesn't
- Cost: Wasted trip for a crew
- False Negative: model predicts a pump is functional when it actually needs repair
- Cost: A community is left without clean water.
- Recall is metric that minimizes False Negatives.

Modeling:

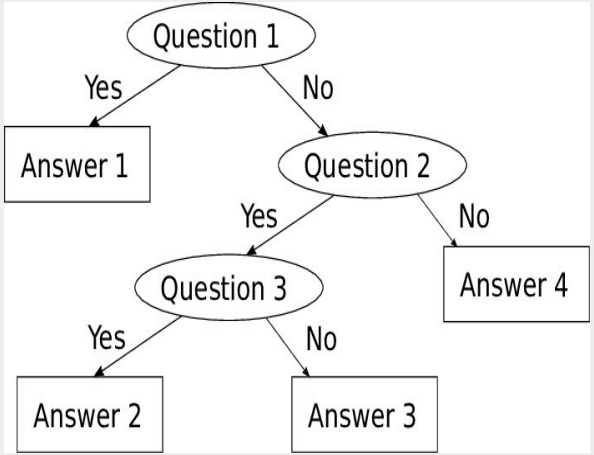
Logistic Regression:

- "yes" or "no" decision.
- Recall Score: 0.66



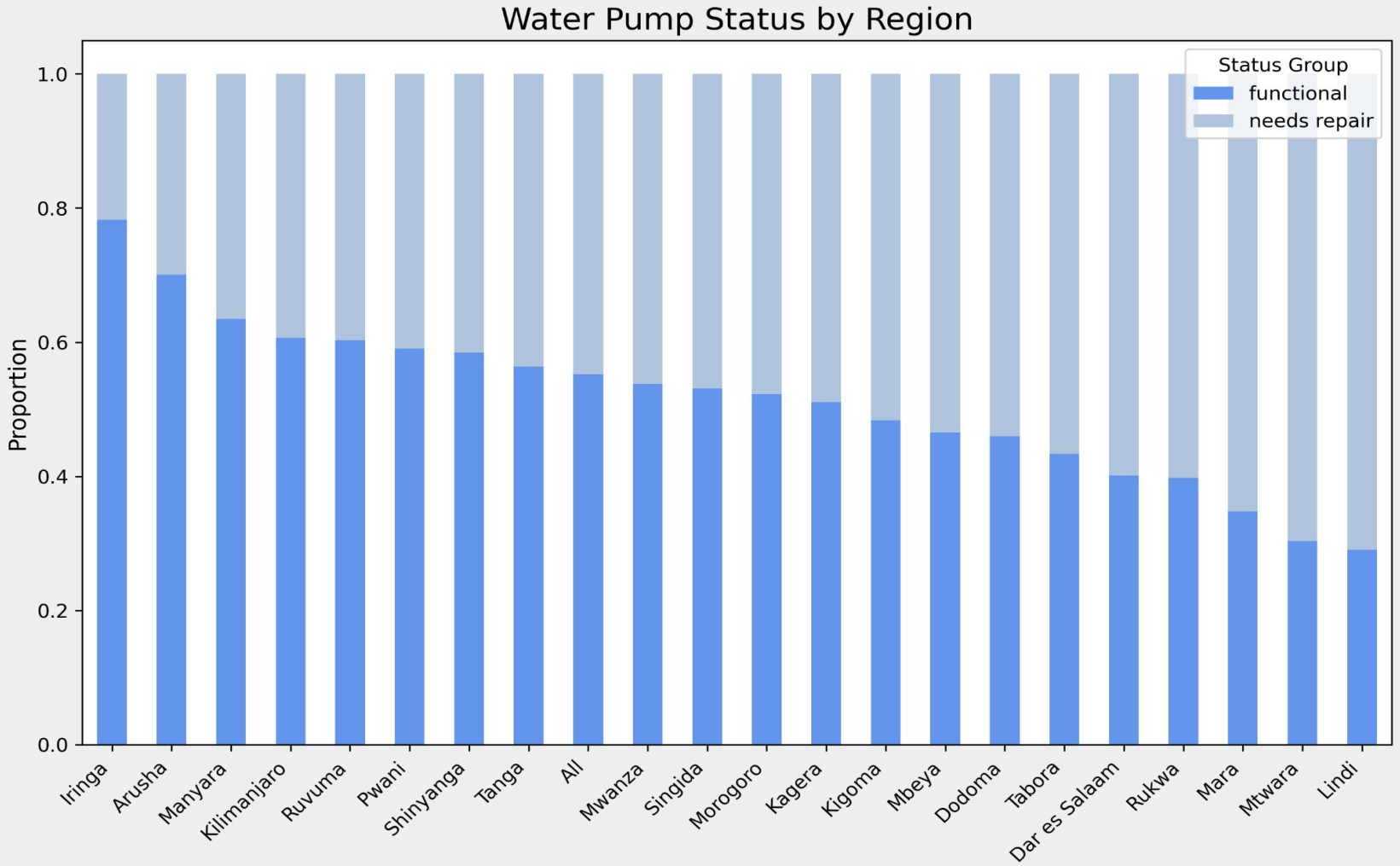
Decision Tree:

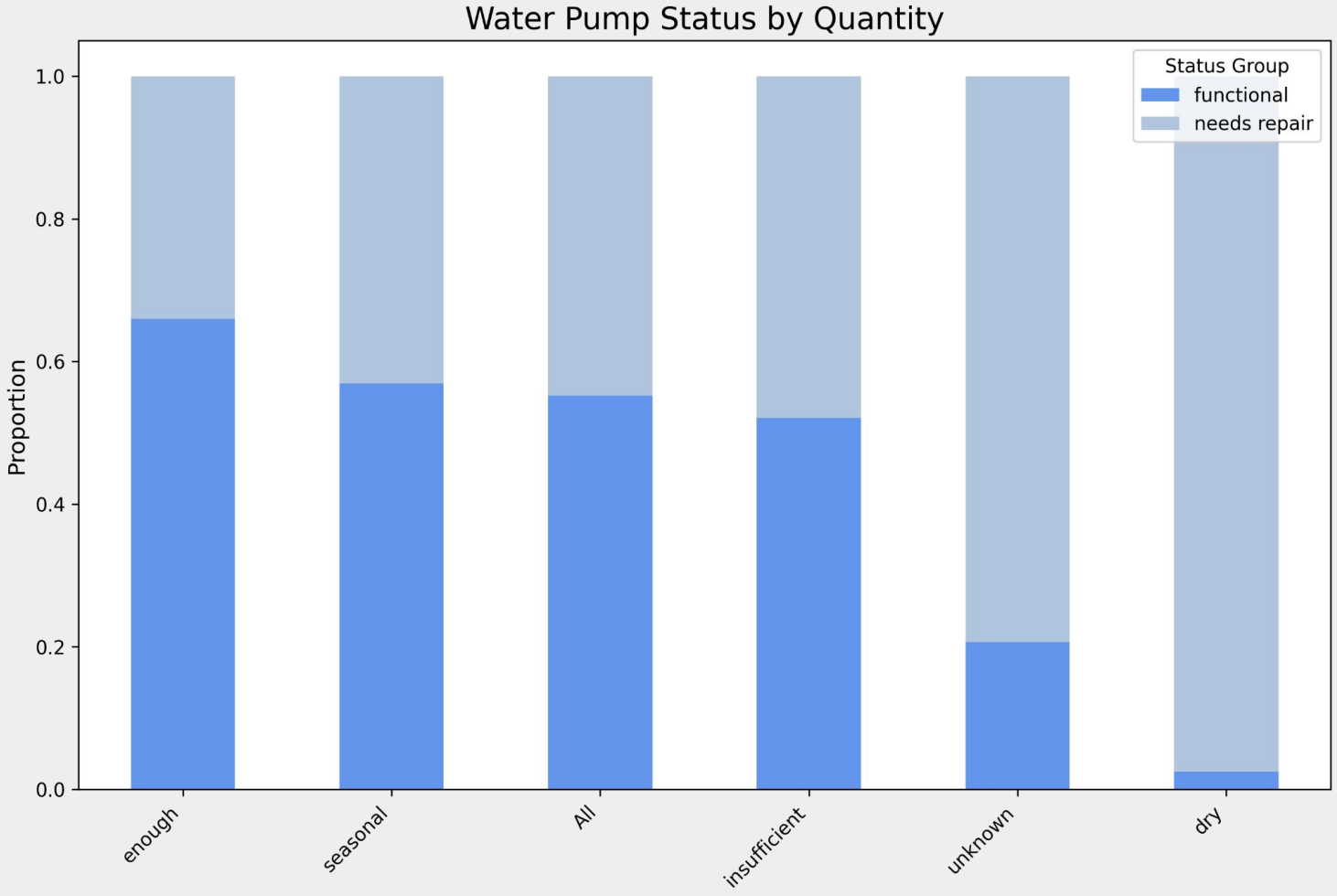
- "if-then-else" decisions
- Recall Score: 0.75



Model Evaluation Metrics:

Classification Metric	Numeric LogReg	Combined LogReg	Tuned and Weighted Combined LogReg	Decision Tree
Accuracy	0.62	0.75	0.73	0.78
Recall (Class 1)	0.48	0.60	0.66	0.75
Precision (Class 1)	0.6	0.79	0.73	0.76
F1-Score (Class 1)	0.53	0.68	0.69	0.75


Water Pump Status by Region
Status Group
functional
needs repair
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Iringa
Arusha
Manyara
Kilimanjaro
Ruvuma
Pwani
Shinyanga
Tanga
All
Mwanza
Singida
Morogoro
Kagera
Kigoma
Mbeya
Dodoma
Tabora
Dar es Salaam
Rukwa
Mara
Mtwara
Lindi


Water Pump Status by Quantity
Status Group
functional
needs repair
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
enough
seasonal
All
insufficient
unknown
dry

Recommendations:

1. **Geographically-Focused Maintenance:**

   Prioritize regions with the highest number of pumps needing repair to maximize impact

2. **Develop a Quality Control Program for Installers:**

   Look at installer performance to prevent future breakdowns from human error

3. **Explore Advanced Models:**

   Use models like Random Forest or Gradient Boosting to push predictive results even higher

Limitations and Next Steps

Current Limitations:

- Data is old (most recent is from 2013)
- Messy data with lots of missing values

Next Steps:

- Improve data collection
  - Gather more recent and more extensive data on water pumps
  - Record maintenance/repair information in real time
- Switch to a proactive system by implementing a predictive model that labels pumps as high-risk before they stop working
- Prioritize maintenance and repairs

Thank you.

Any Questions?

Morgan Nash

morganmichellenash@gmail.com