H2knOw

Using AI to connect people to safe water

Team

Anuradha (Annie) Passan

Ethan Duncan

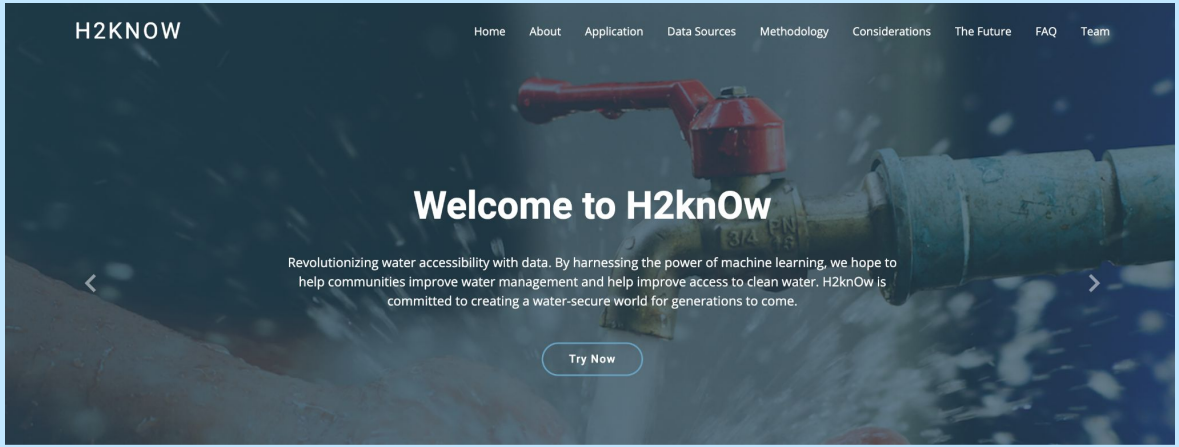
Ivan Escalona

Bailey Kobuke

Motivation

- **Ensure safe water access for all**
  - UN 2030 Agenda for Sustainable Development
- Obstacles:
  - Failure to meet increasing demand
  - Climate change
  - Lack of monitoring systems
- By 2025, 66% of the global population will experience water stress, i.e. face difficulties in accessing safe water (UNDESA)

H2knOw



H2kn**O**w is a web interface allows **individuals** in water stressed countries to locate the closest functioning water points to their location.

H2KnOw

Proof of Concept

Currently a web interface, but would eventually add a mobile app version

Users

End users are individuals looking for safe water; would be jointly administered with local partners (govt., NGOs)

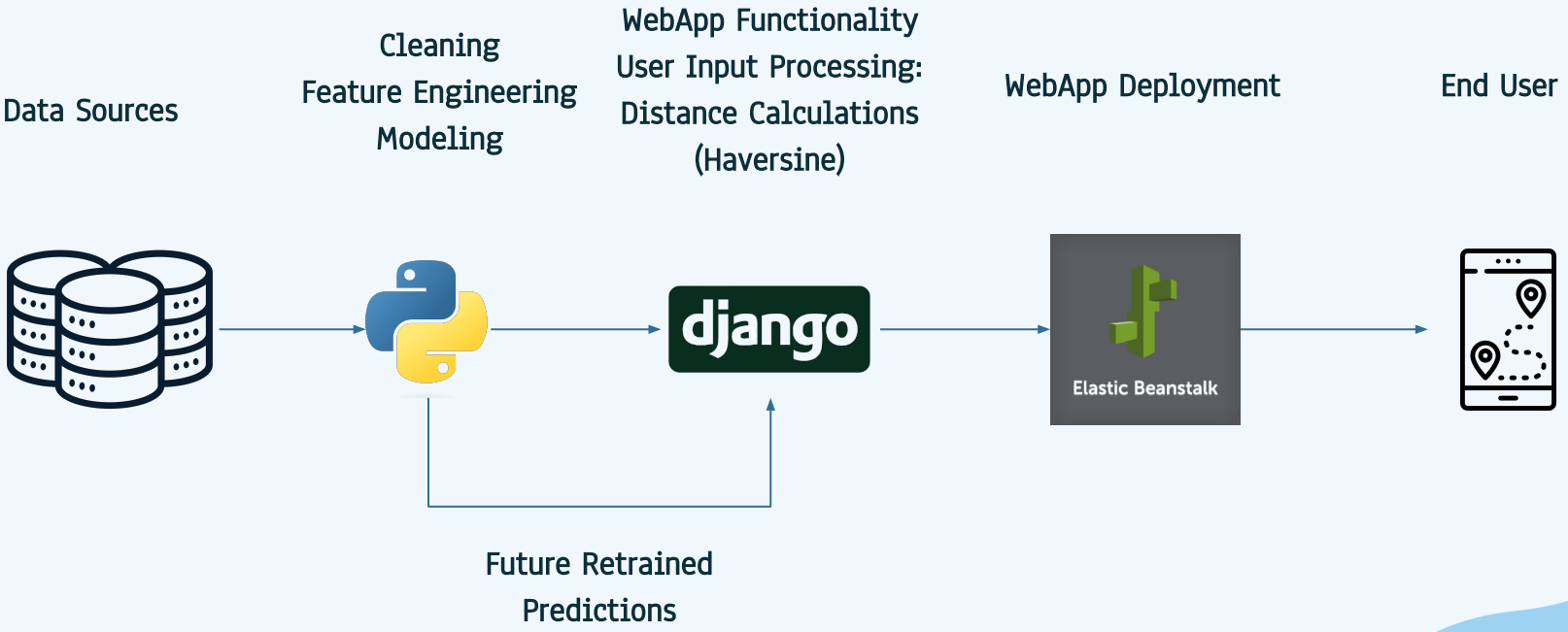
Countries

Sierra Leone, Nigeria, and Uganda

Market Size: 136 M

Based on joint population using the internet as of Jan. 2023

H2knOw Workflow



Data

- Platform for sharing **water point data** collected by NGOs, government agencies, and local communities
- **Data:** functionality, location, management, water quality, seasonality, proximity to roads & towns

- **Country** development indicators
- **Data:** GDP growth, agricultural development, weather, political stability, rule of law

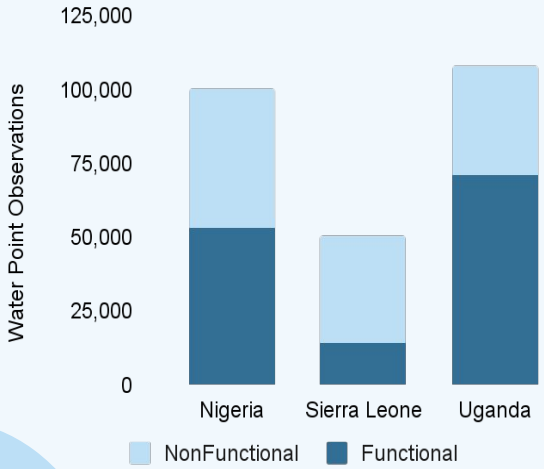
- **Region-level** development indicators
- **Data:** GNI per capita, human development index, agricultural employment, avg. year female schooling

Data Extraction & Preparation

1. Data Extraction
2. Cleaning
3. Feature Engineering
4. Data Join Secondary Data

EDA Highlights

Water Point Status



259,518

Observations

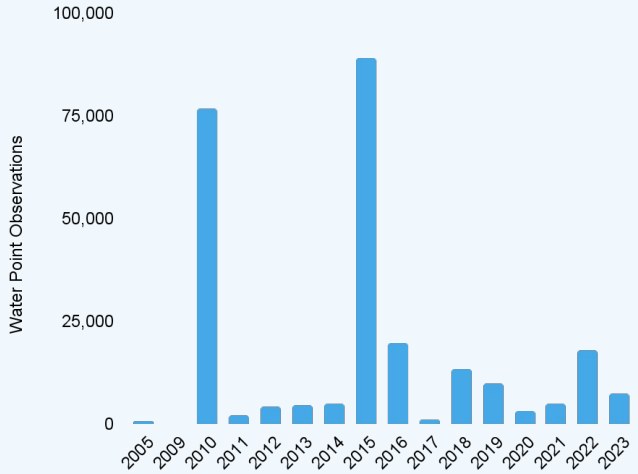
53%

Functional

47%

Non Functional

Observation Cadance



Time Series Classification

Observations

Many water points have historical data on their functionality

Time Attributes

Observation month and day useful to extract annual and monthly trends relating to functionality

COVID-19

Pandemic changed population behavior; affect TS models.
Highest # cases in 2021, with impact seen in 2022-2023 observations.
Decision: use observations until 2021.

XGBoost Classification

XGBoost: a library that runs gradient boosting, which is an ensemble ML algorithm that combines several weak decision tree learners sequentially to form a strong learner.

Why XGBoost?

- Ability to handle nulls
- Scalability (parallel processing)
- Robust to multicollinearity
- High accuracy
- Built-in regularization

Model Pipeline

1. **Data Pipeline**
   54 features
2. **Data split**
   Train: 2005 - 2020 [229,230 obs.]
   Test: 2021 [4,965 obs.]
3. **Binary Encoding**
4. **Training**
   CV time series split (5 fold)
5. **Test & Evaluation**

Modeling Results

1. Random Guess

- Metrics:
  - **Accuracy:** 58%
  - **F1 Score:** 77%

2. Decision Tree

- Metrics:
  - **Accuracy:** 74%
  - **F1 Score:** 2%

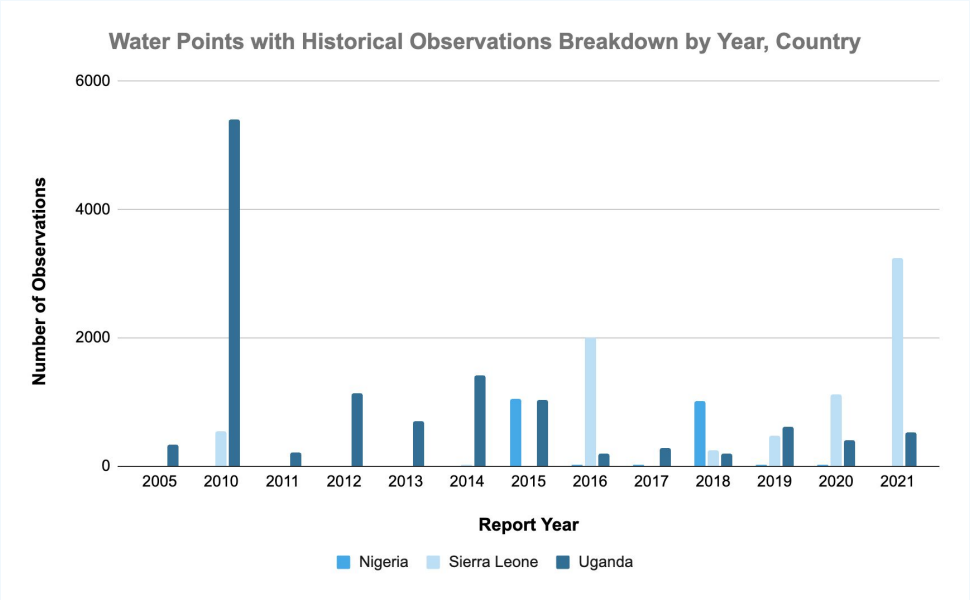
Baseline Models

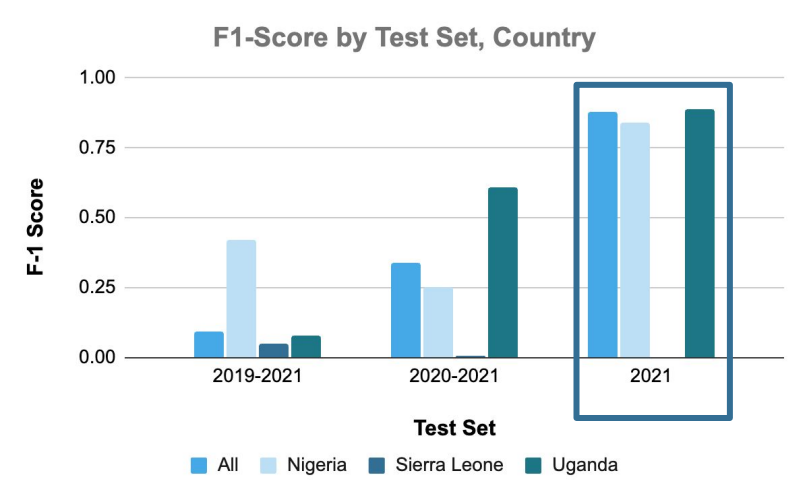
3. XGBoost (Time-Series)

- Metrics:
  - **Accuracy:** 98%
  - **F1 Score:** 89%
- Top features: **staleness, lon, lat, install year, roads & city proximity, pressure score, crucialness, log GNI per capita**
- Predictive capability strong, but will never be perfect

Best Model

Further Evaluation:
Time Spilt, Country Fairness





Key Takeaways:

- Training data distribution by class & year affect learning
- More data needed to take to production

Demo

Demo Example

Search Point (Sierra Leone): 9.693440, -12.266625

- The closest waterpoint (6CX9MPVM+437) is predicted to be nonfunctional, it is omitted from the map
- The 3 closest water points predicted to be functioning are:
  - 6CX9MPRM+X44
  - 6CX9MPWJ+H9P
  - 6CX9MQJG+49X

Model Output

	wpdx_id	predictions	distance
1	6CX9MPVM+437	0	0.107071
2	6CX9MPRM+X44	1	0.131460
3	6CX9MPWJ+H9P	1	0.444762
40	6CX9MQHC+4V4	0	4.600884
5	6CX9MQJG+49X	1	4.879108
11	6CX9MQJG+29M	0	4.895330
35	6CX9MQFF+6FJ	1	4.959030
20	6CX9MQFF+PR5	0	4.996269
22	6CX9MQFF+4H2	1	4.999265

Moving to Production

Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
Partnerships	**Monitor**	**Retrain**	**Rollout**	**Grow**
Engage with actors working with locals to help rollout and collect more water point and granular context data	Monitor data for pre-pandemic behavior	Retrain model with newer data	Start at district or region level before expansion	Model maintenance, Add countries, Incorporate feedback systems, Mobile app

wpdx water point data exchange

unicef

WASH Cluster Water Sanitation Hygiene

Govts.

Conclusion

Reliably accessing safe water is increasingly becoming difficult.

H2KnOw can help by connecting locals to safe water points.

Moving forward, we would first look to developing partnerships with relevant actors to take this to production.

Thank you

User Accessibility

	Population (total: 281)	2023 Internet Penetration	# Mobile Connections	Smartphone user forecast
Sierra Leone	8.7 M	21%	113%*	18% growth 2023-28
Nigeria	223.8 M	55.4%	87.7%	>140 M by 2025
Uganda	48.5 M	24.6%	63.8%	10 M in 2022; growing

Key Takeaways:

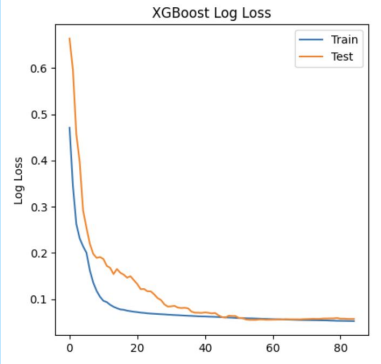
1. There is space for growth in terms of connectivity in all three countries.
2. But as connectivity continues to improve, and more individuals adopt smartphones, the accessibility of H2KnOw as a web interface & mobile app will increase.

* Individuals can have >1 mobile connections

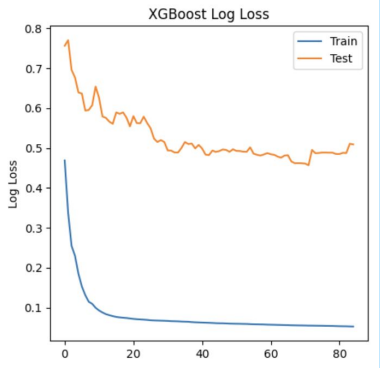
Competitive Landscape

Company	Product / Solution overview	Primary customers	H2KnOw differentiation
WPdx [Non-profit]	Water point functionality	NGOs, governments, international organizations	H2KnOw used XGBoost, vs Light GBM; different contextual data; H2Know achieved lower loss score (~0.1) compared to WPdx (~0.3) Also H2KnOw's end user is individual, providing closest functioning waterpoints; WPdx is simply a map of predicted functionalities.
Akvo [Enterprise]	Data-driven decision making and innovative technology to the development sector	NGOs, governments, international organizations	H2KnoW focuses specifically on water point functionality, while Akvo focuses more on water management.
Gybe [Startup]	Satellite imagery targeting water insecurity	Governments	H2KnOw focused on feature based detection vs image based detection to predict water quality.

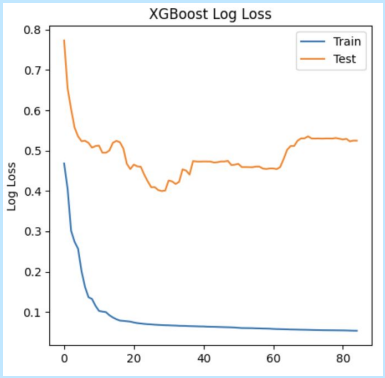
Loss Curves by Time Split



2021



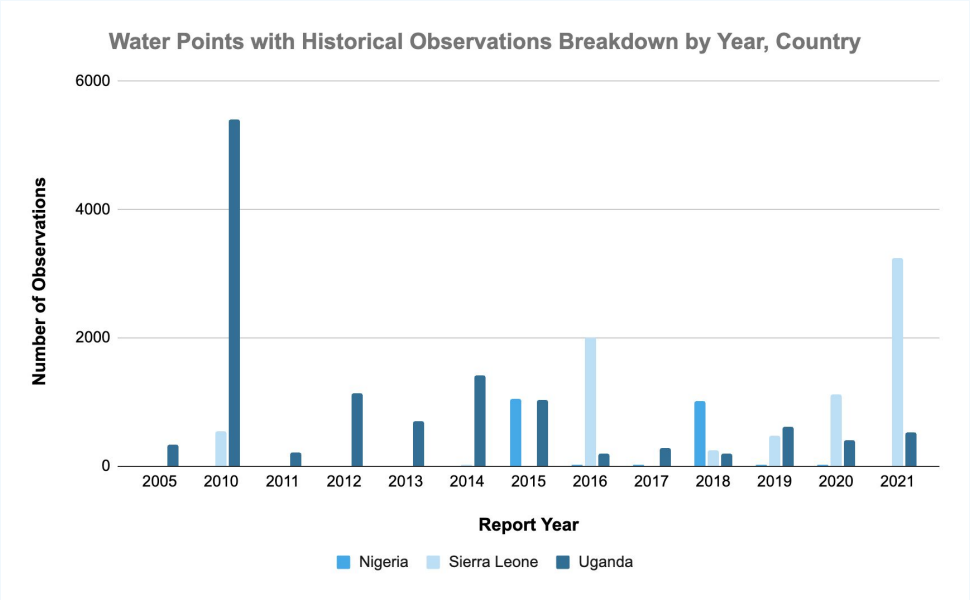
2020-2021

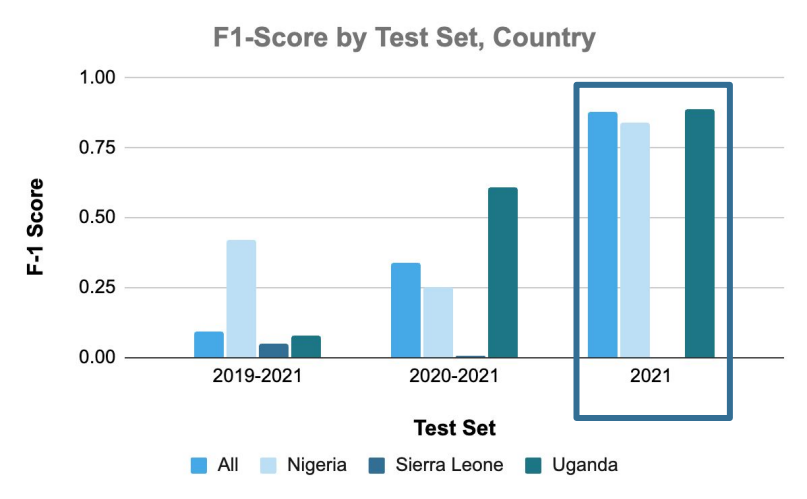


2019-2021

The model is not able to generalize well and finish learning when the training does not go until 2020.

Further Evaluation:
Time Spilt, Country Fairness





Key Takeaways:

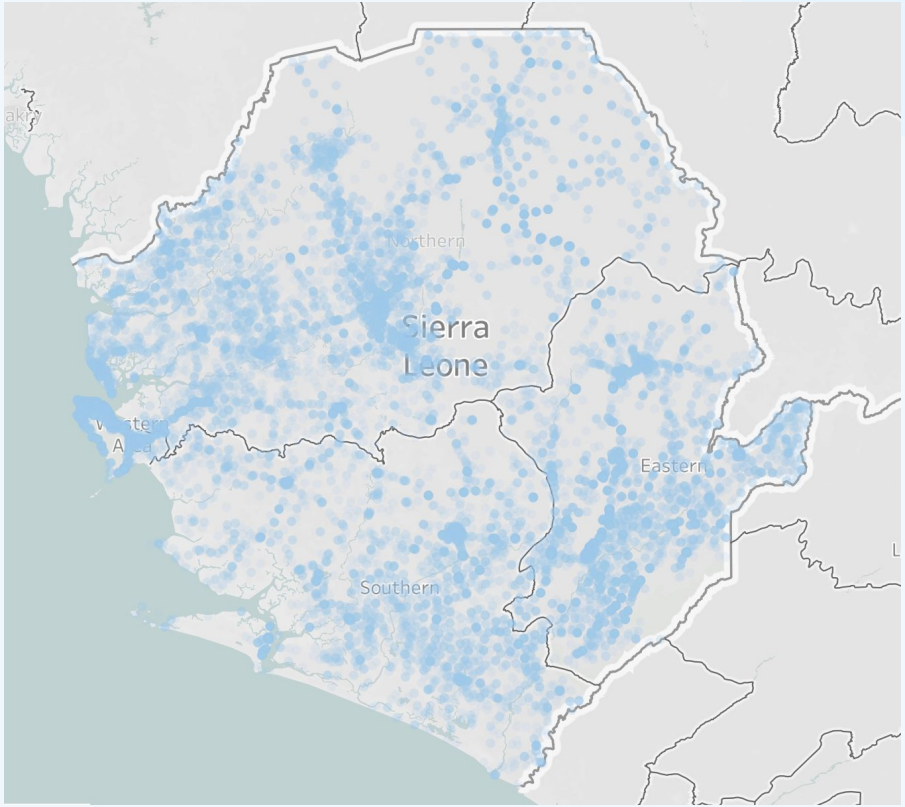
- Training data distribution by class & year affect learning!!
- Model degrades with larger test sets
  - Likely due to class imbalance for countries by year
  - Hinders model's ability to learn and generalize
- Better with Uganda & Nigeria, worse with Sierra Leone
  - Uganda has most spread out historical data
  - Sierra Leone has more recent data spilling to test sets
- More data needed to take to production

Sierra Leone Water Points

Observations: 27,011

Functional: 13,181

Non-Functional: 13,830

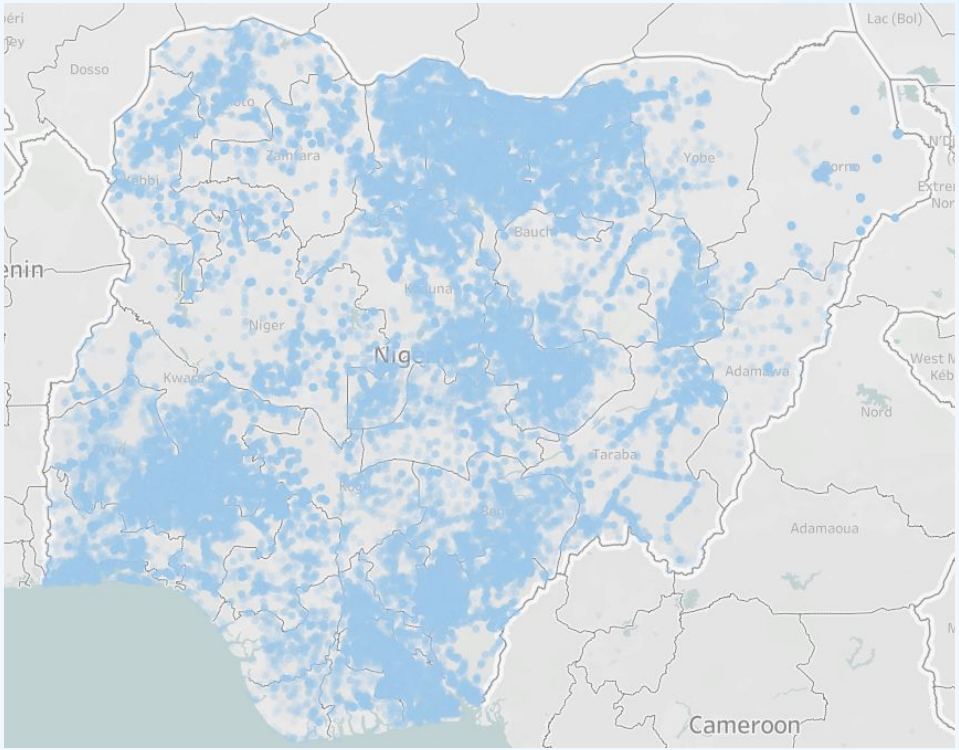


Nigeria Water Points

Observations: 95,210

Functional: 53,031

Non-Functional: 42,179

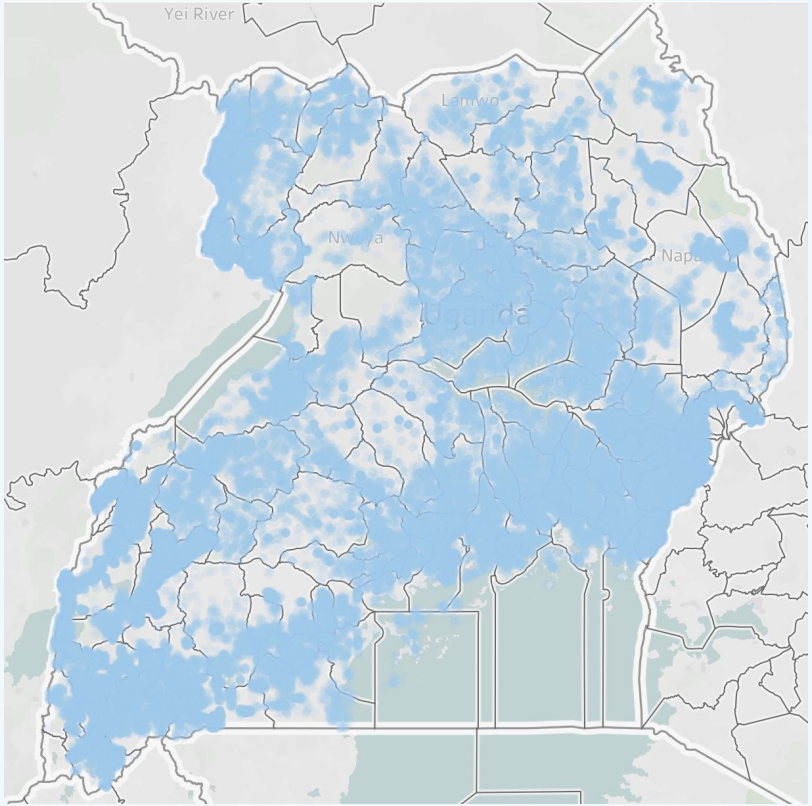


Uganda Water Points

Observations: 92,529

Functional: 65,594

Non-Functional: 26,935



Key Challenges

Data Timeframe

The WPdx dataset, originally, only has a single observation per waterpoint

Data Inconsistency

Given the nature of data collection and maintenance, our data has built in difficulties i.e. high class imbalance in 2021 versus average in dataset

Experience Gaps

Our team had limited web dev experience, requiring extensive trial & error to troubleshoot any site issues

Hyperparameter Tuning

list of hyperparameters tuned

- N_estimators
- Min_child_weight
- Gamma
- Subsample
- Colsample_bytree
- learning_rate

final model hyperparameters

- N_estimators: **85**
- Min_child_weight: **1**
- Gamma: **0**
- Subsample: **0.9**
- Colsample_bytree**: 0.7**
- Learning_rate**: 0.1**