Identifying Broken & Non-functioning Water-points in Tanzania

MARK PATTERSON, DECEMBER 2020

43% of people in Tanzania…lack access to a reliable, safe water source

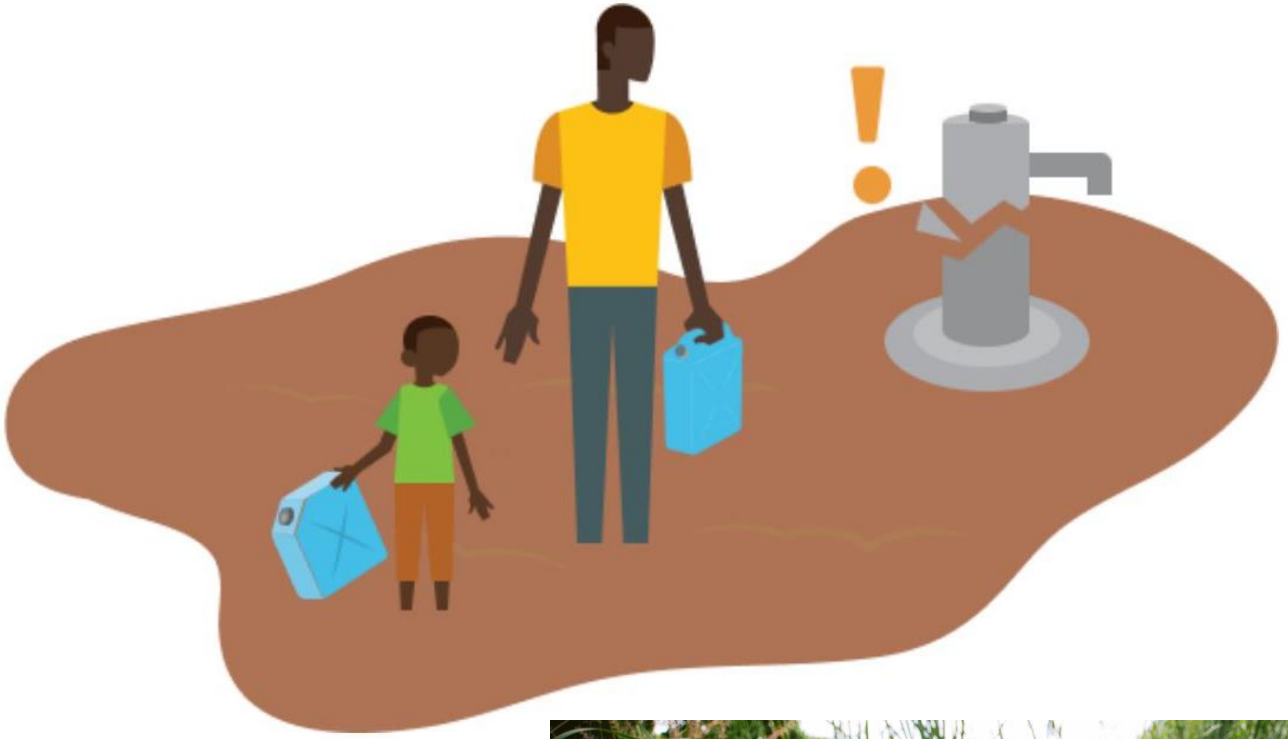
- Drinking
- Cooking
- Washing
- Growing plants / gardens

Meet Annah:

Lack of a clean water source means:

- Limited food options
- Poor sanitation Disease risk
- Exhaustion - long walk each day
- Risk of violence and rape
- Limited time at school

Wells, pumps, and storage often in disrepair



GOAL:

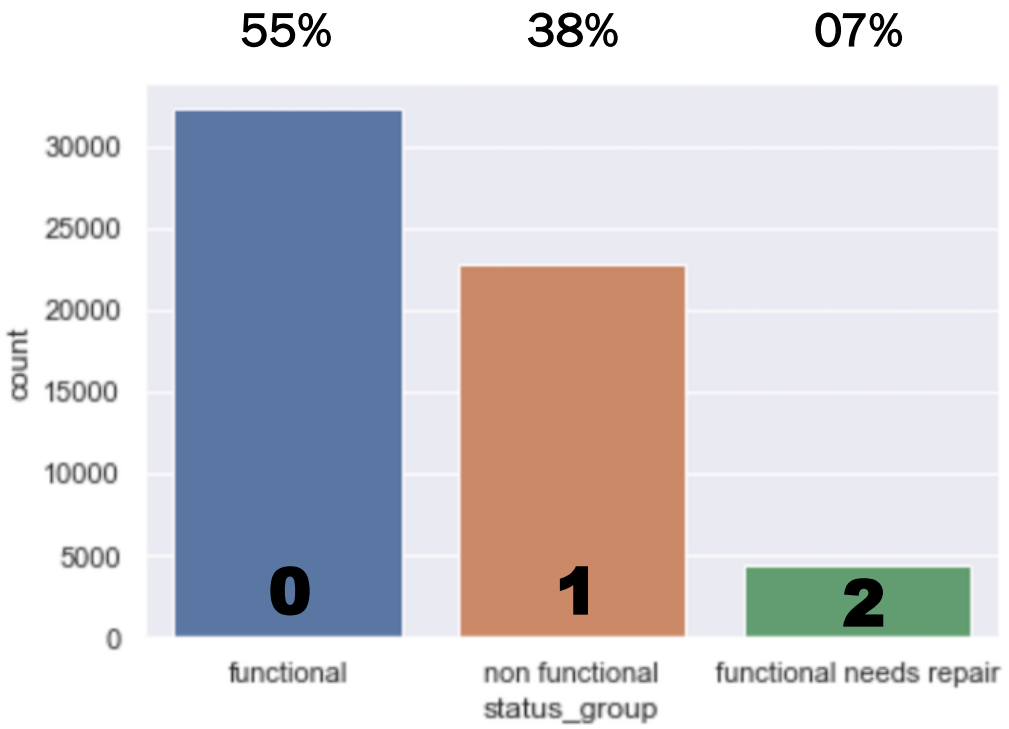
A supervised machine learning model to classify water-points to identify and predict which ones need rehabilitation or repair

Data Source:

Tanzanian Ministry of Water

TAARIFA (tech non-profit)

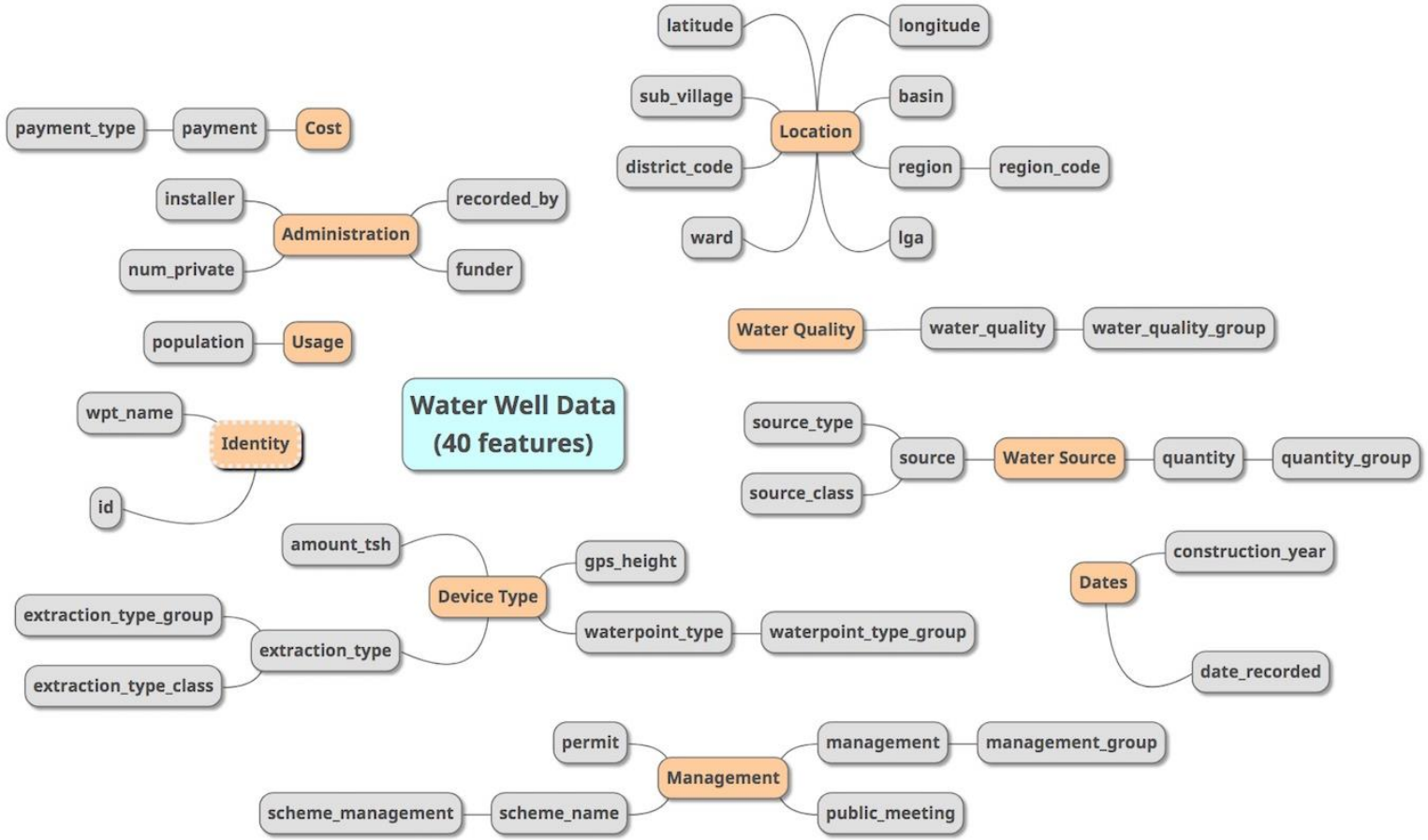
- Waterpoints: **1960 – 2013**
- **59,400** records
- **41** variables



Data Analysis & Shaping

Top Issues :

- Redundant variables dropped
- Large number of 0 values (left, transformed, dropped)
- Large number or “other” or “unknown”



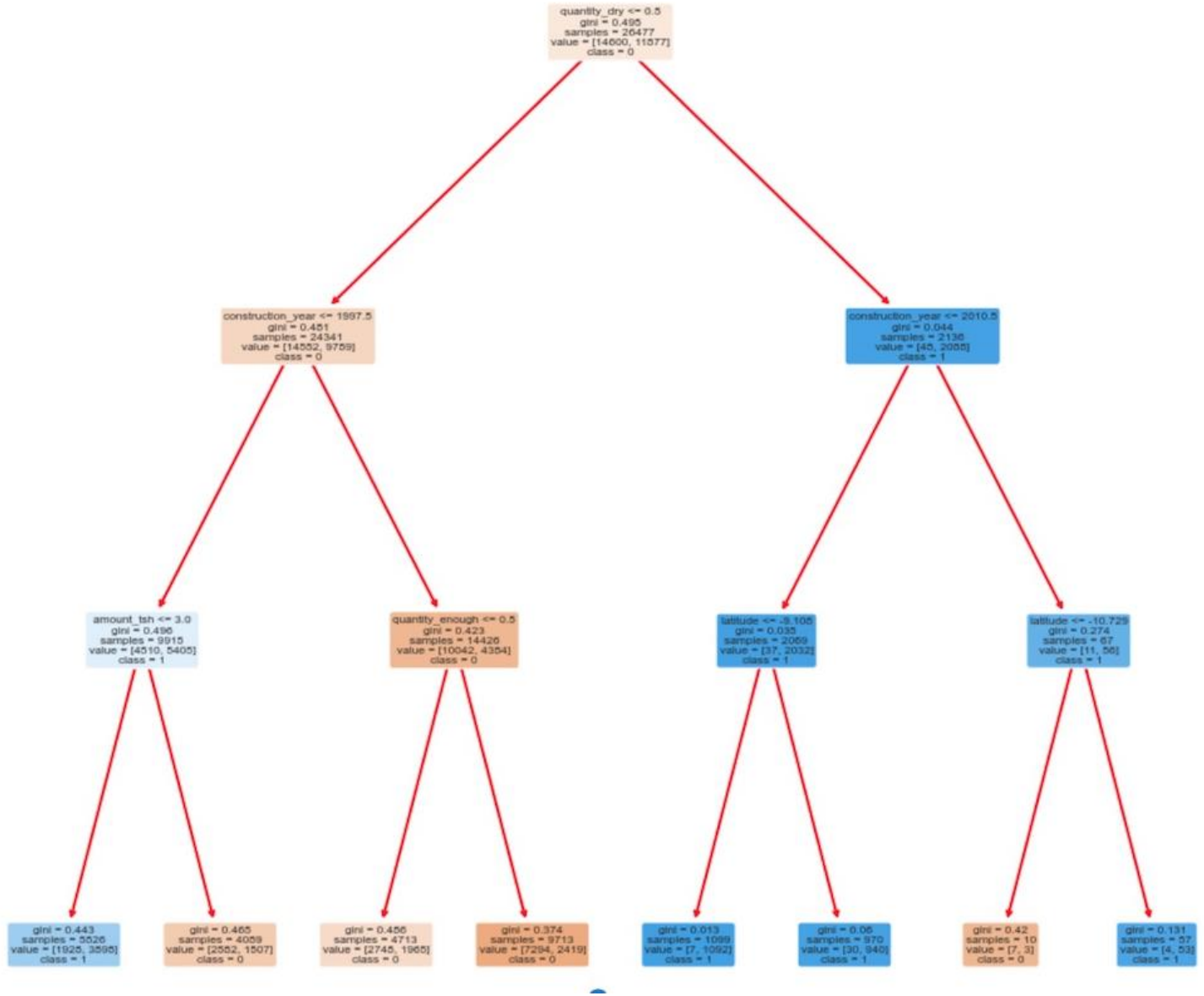
59,400 records
41 variables

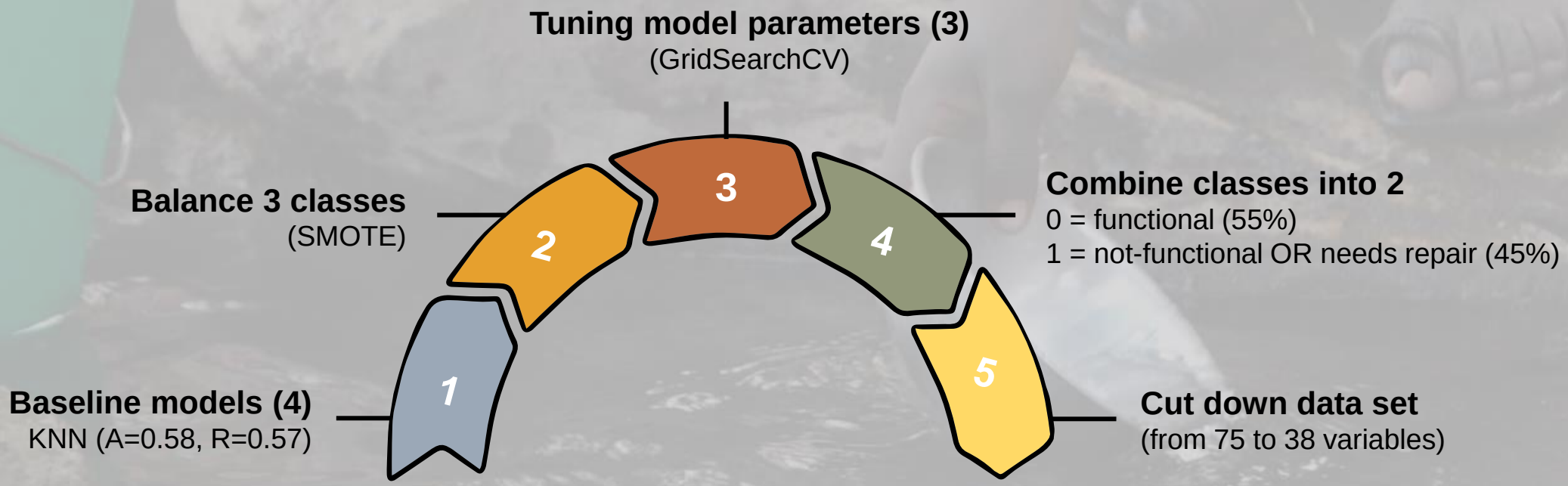
35,303 records
17 variables

35,303 records
75 variables

Classification Models Applied

- KNN
- Decision Trees
- Random Forest
- XGBoost





The Iterative Modeling Process

A total of 24 models were run via this iterative refinement process.

Combining target classes into 2 loses some granularity but creates fairly balanced classes and makes conceptual sense.

```
--------------------------------------------------------------------------
--------------------------------------------------------------------------
Classification Report - TEST  - XGBoost-C-Optimized2
--------------------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.82      0.87      0.84      4839
           1       0.82      0.77      0.79      3987

    accuracy                           0.82      8826
   macro avg       0.82      0.82      0.82      8826
weighted avg       0.82      0.82      0.82      8826

--------------------------------------------------------------------------
Confusion Matrix - TEST - XGBoost-C-Optimized2
--------------------------------------------------------------------------
Predicted     0     1   All
True
0          4186   653  4839
1           926  3061  3987
All        5112  3714  8826
--------------------------------------------------------------------------
```

True Positives for class 1

Best Model: XGBoost (2-class, reduced data)

- 82% of the time this model should be able to accurately predict if a waterpoint is functional or not.
- With recall of 0.77 for class 1, there is still a 23% chance of a false negative (it's not working but predict it is)

What factors contribute most?

Important to models:

- Age of water-point
- Quantity of water
- Location

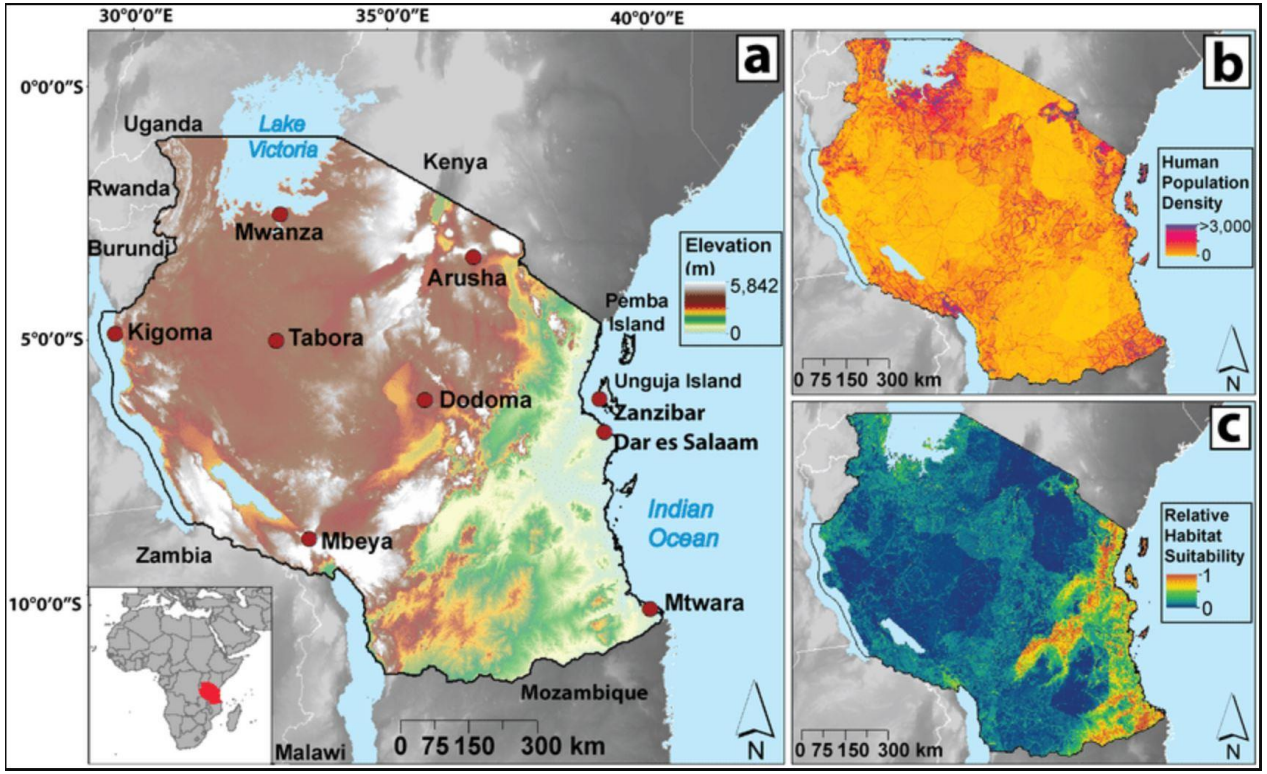
	features	importance
0	quantity_dry	0.154350
1	longitude	0.137392
2	waterpoint_type_group_other	0.130257
3	latitude	0.107340
4	construction_year	0.102898
5	gps_height	0.072590
6	population	0.039401
7	amount_tsh	0.039079
8	quantity_enough	0.025860

In Conclusion

- Model could help predict water-points in need of service
- Help prioritize resources, funding, training
- Provide Annah and others sustainable access to safe, clean water

Next Steps

- Use location data to visualize water-points in relation to geography, climate, population
- Explore various approaches to data preparation and shaping
- Refine modeling approach to use pipelines and functions



Thanks to…

Flatiron School

- Our instructor – Yish
- Staff –
- Fellow students

Tanzanian government, it's citizens, NGOs, volunteers and donors who continue to address this serious issue…

May we soon have safe water for all!