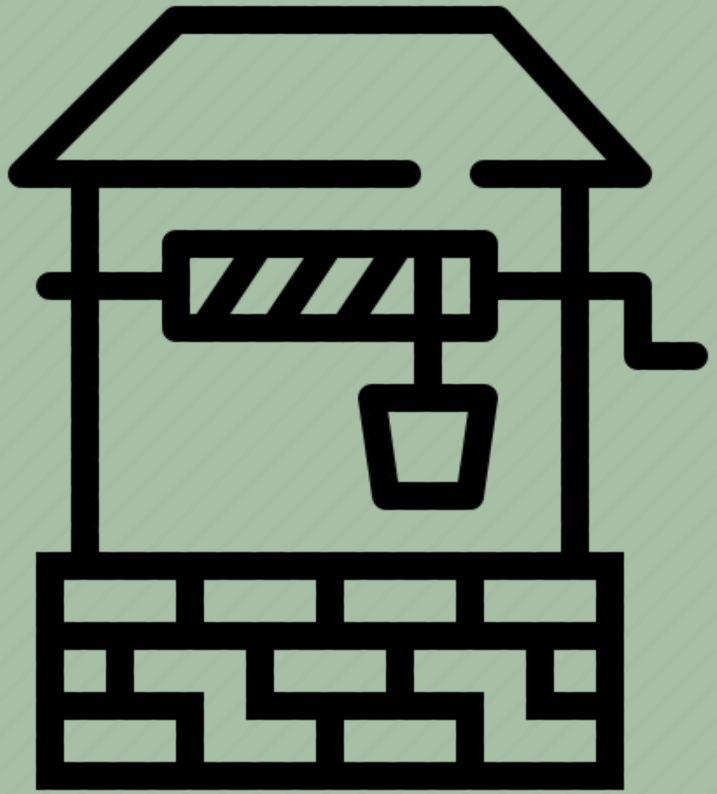
Tanzanian Water Well Analysis

Aaron Cherry, Victor Chen



Model Prediction of Water Point Condition

- Tanzanian Ministry of Water
- Enhance maintenance operations
- Ensure clean water to population

Data Understanding

Dataset

Taarifa Waterpoints Dashboard

Features

Age of waterpoint
Location
Usage
Quality of water

Datapoints

57,588 data points

Target

31,389
54.51%

Functional

3,931
6.82%

Functional, needs repairs

22,268
38.67%

Non-functional

Key Features

Water Quantity

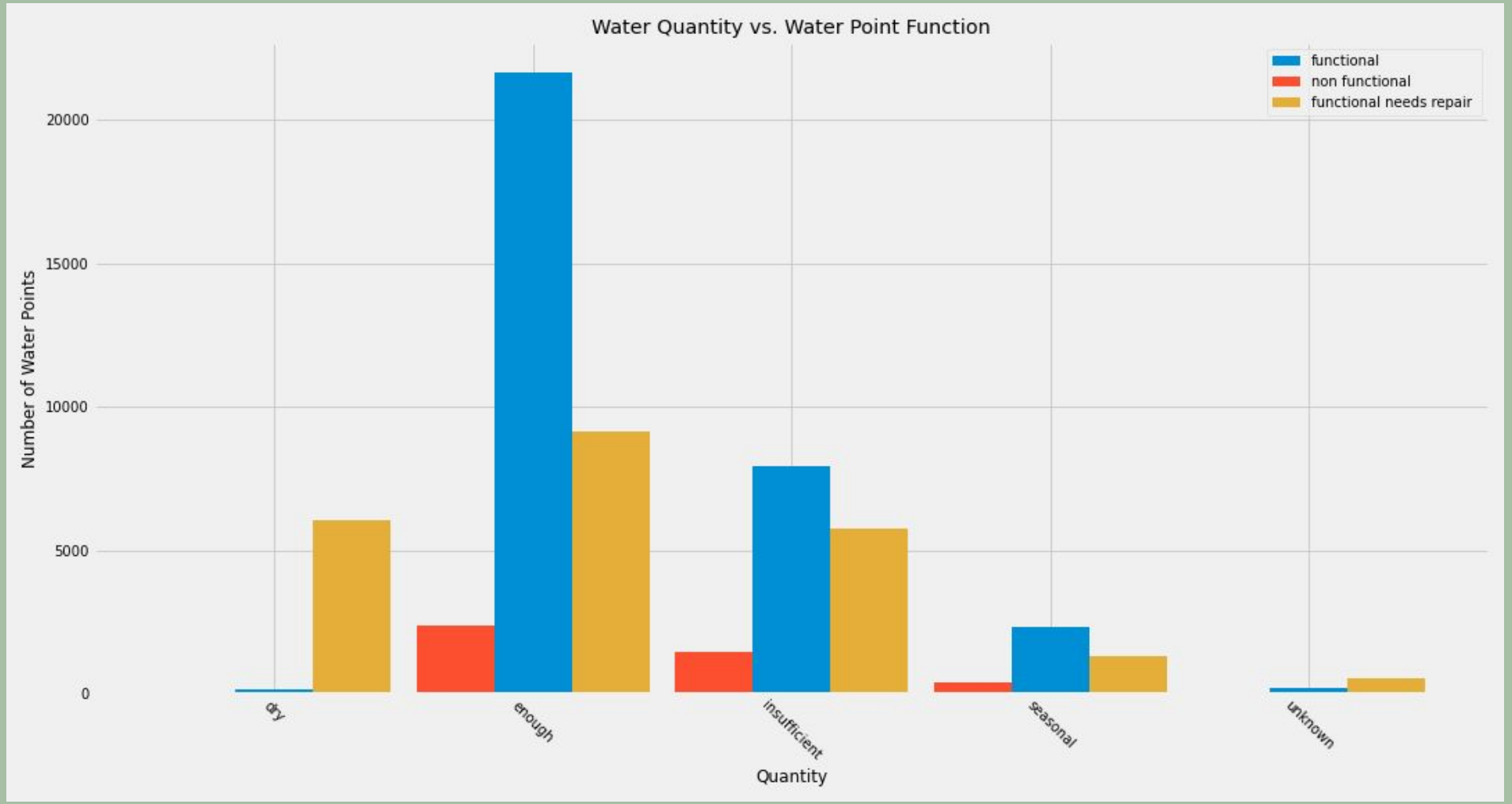
Amount of water the point serves

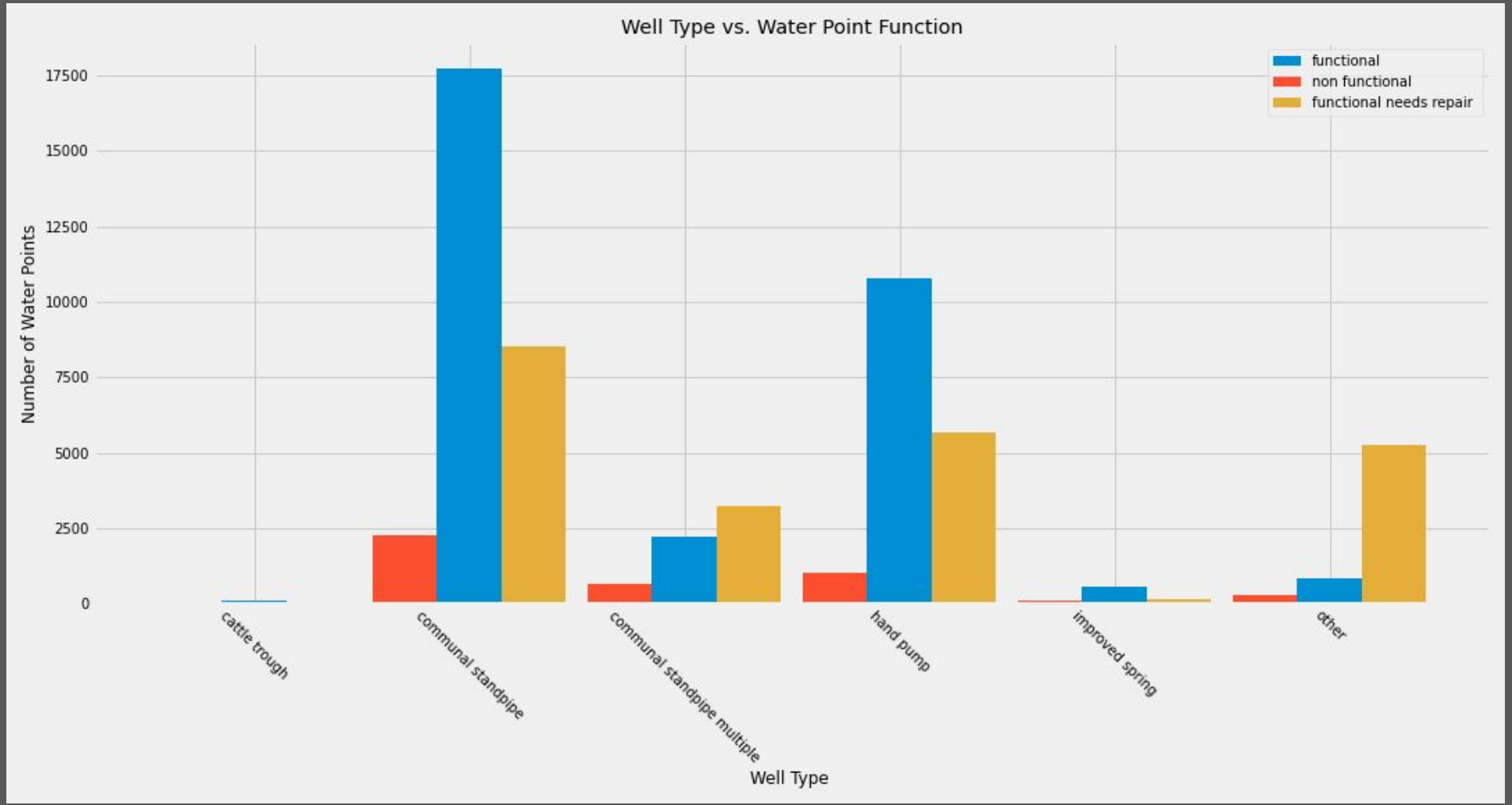
Type

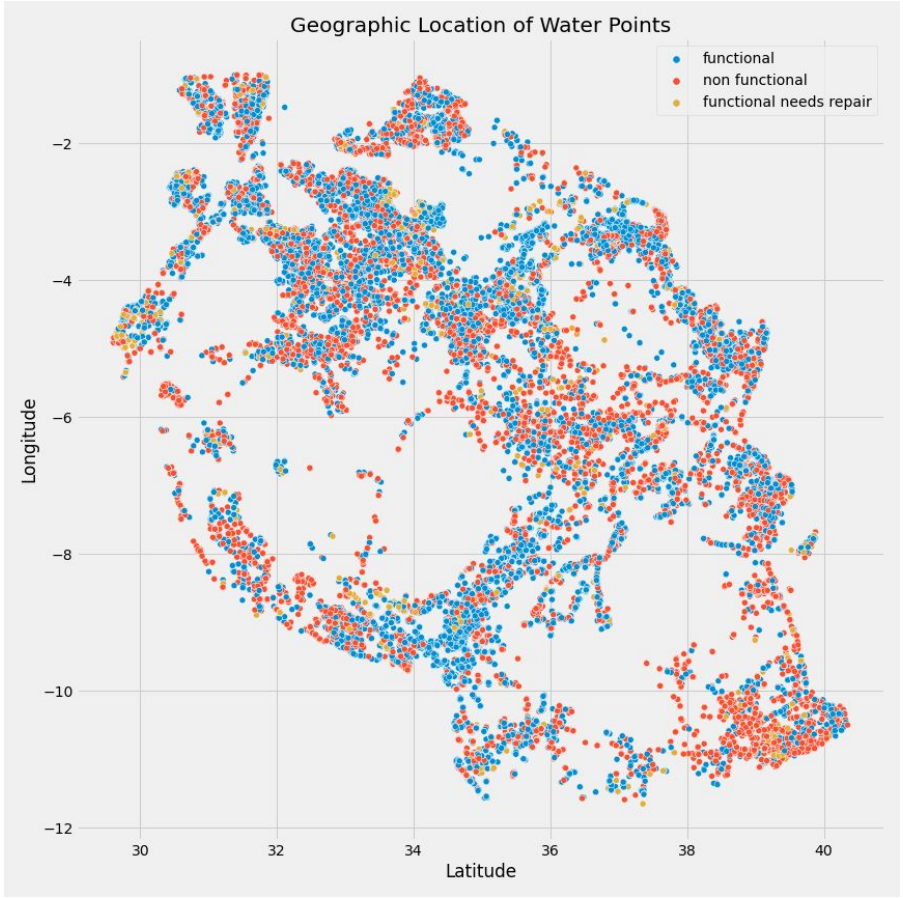
Type of water point

Location

Geographic Location of the water point


Water Quantity vs. Water Point Function
functional
non functional
functional needs repair
Number of Water Points
20000
15000
10000
5000
0
dry
enough
insufficient
seasonal
unknown
Quantity


Well Type vs. Water Point Function
functional
non functional
functional needs repair
Number of Water Points
17500
15000
12500
10000
7500
5000
2500
0
cattle trough
communal standpipe
communal standpipe multiple
hand pump
improved spring
other
Well Type


Geographic Location of Water Points
functional
non functional
functional needs repair
Longitude
Latitude
−2
−4
−6
−8
−10
−12
30
32
34
36
38
40


Nairobi
Rwanda
Burundi
Mwanza
Arusha
Malindi
Kilifi
Mombasa
Ujiji
Dodoma
Zanzibar
Tanzania
Dar es Salaam
Morogoro
Iringa
Mbeya
Kasama
Songea
Mtwara
Mzuzu

Model Metrics

Accuracy

Best predictions for all target classes

Precision

Limit false positives relative to the “nonfunctional” label

Baseline Models

Logistic Regression

Precision: 0.60
Accuracy: 0.67

Decision Tree

Precision: 0.64
Accuracy: 0.75

KNN

Precision: 0.69
Accuracy: 0.77

Random Forest

Precision: 0.68
Accuracy: 0.78

Best Model: Random Forest

- Easy to understand
- Relatively Computationally Efficient
- Simple Optimization Parameters

Random Forest

Precision: 0.68
Accuracy: 0.78

Random Forest Model Tuning

	Precision	Accuracy
Baseline	0.67	0.78
Middle Iterations	0.68	0.79
Final	0.69	0.79

Final Model

Accuracy

0.79

Precision

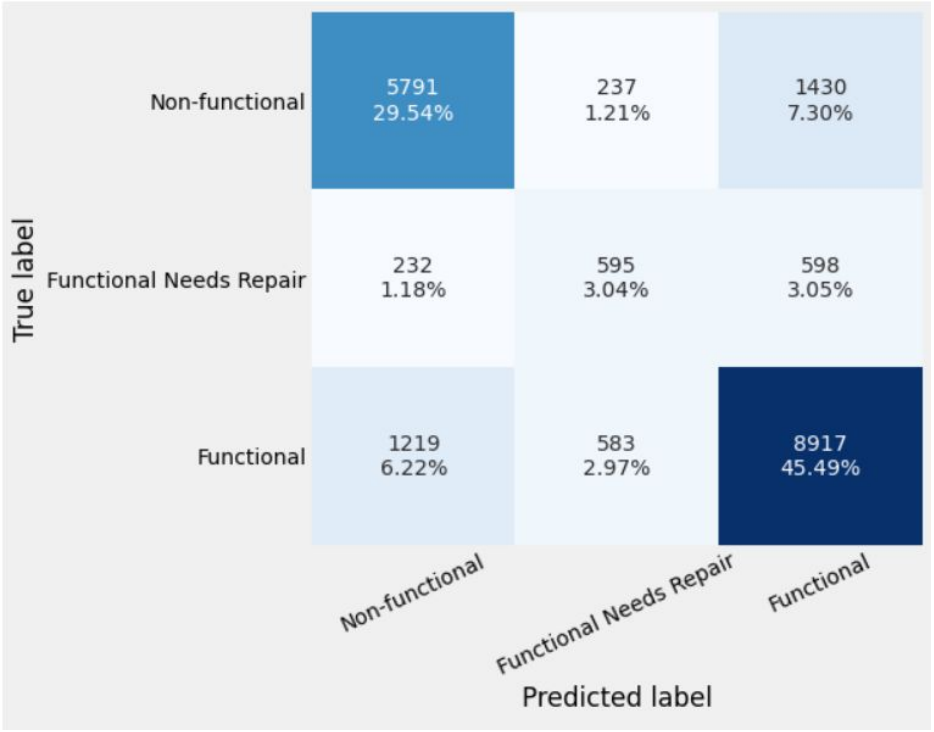
0.69

Parameters

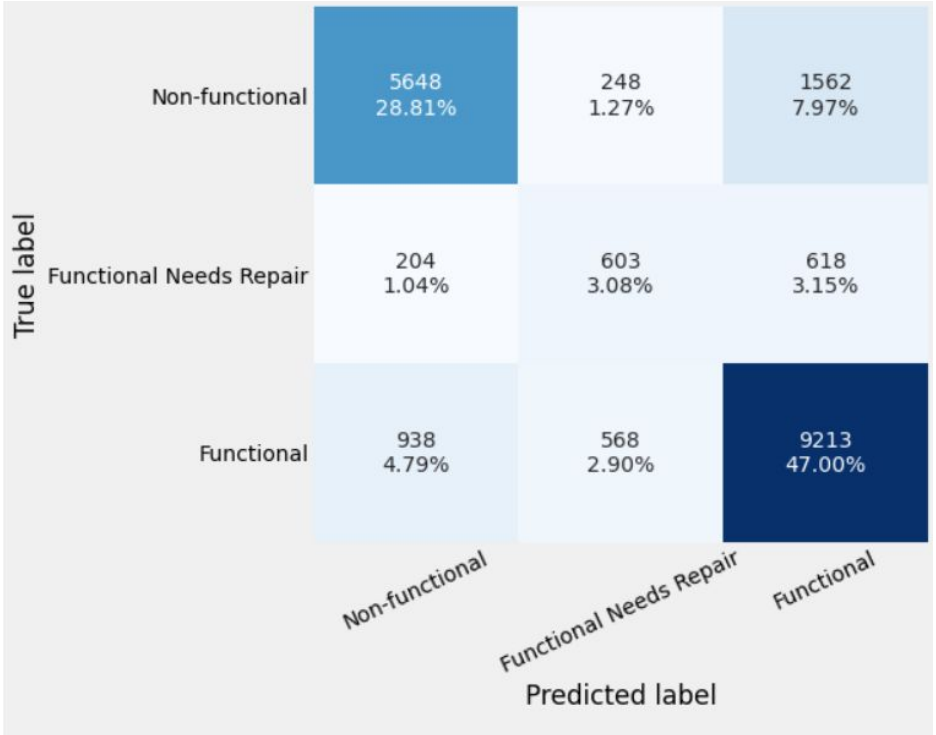
criterion = 'gini'
min_samples_split = 6
n_estimators = 400
min_samples_leaf= 1
max_features = 'auto'

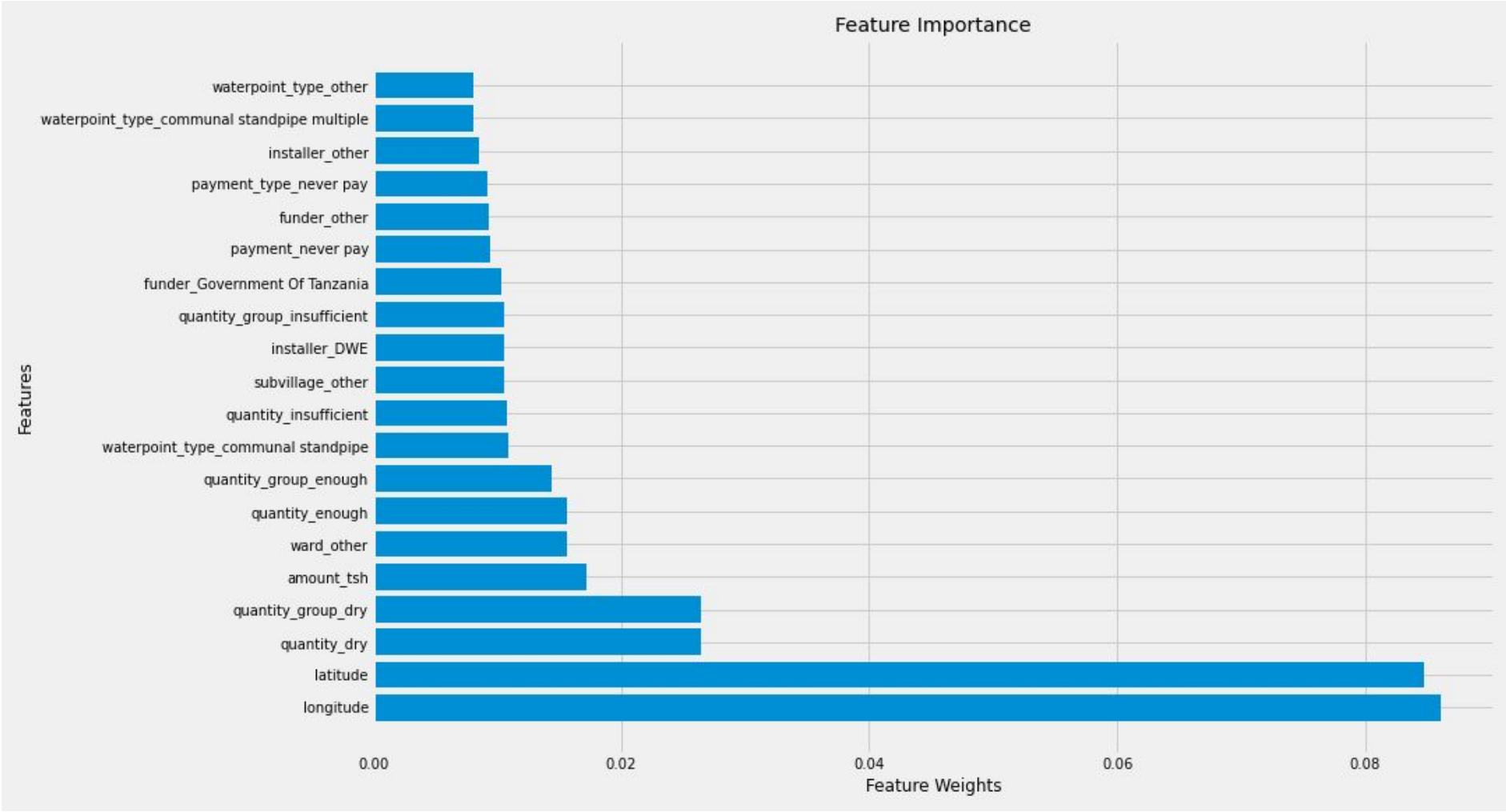
Confusion Matrices

Baseline



Final




Feature Importance
waterpoint_type_other
waterpoint_type_communal standpipe multiple
installer_other
payment_type_never pay
funder_other
payment_never pay
funder_Government Of Tanzania
quantity_group_insufficient
installer_DWE
subvillage_other
quantity_insufficient
waterpoint_type_communal standpipe
quantity_group_enough
quantity_enough
ward_other
amount_tsh
quantity_group_dry
quantity_dry
latitude
longitude
Features
0.00
0.02
0.04
0.06
0.08
Feature Weights

Conclusion

Random Forest

Best Model

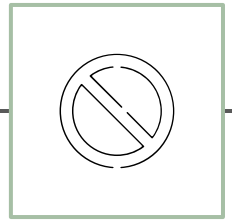
Location, Water Quantity, Waterpoint type

Recommended Features

Next Steps

Features

Consider including features outside of our dataset (latest repair, failure mode)

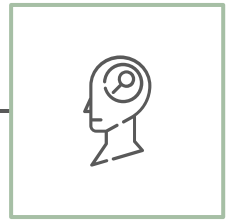


Limitations

Missing location and construction year data for large portion of data.

Updated Data

Consider more recent data to predict failure trend



Model Updates

Predict failure modes
Update model to predict when failures could occur

Thanks!

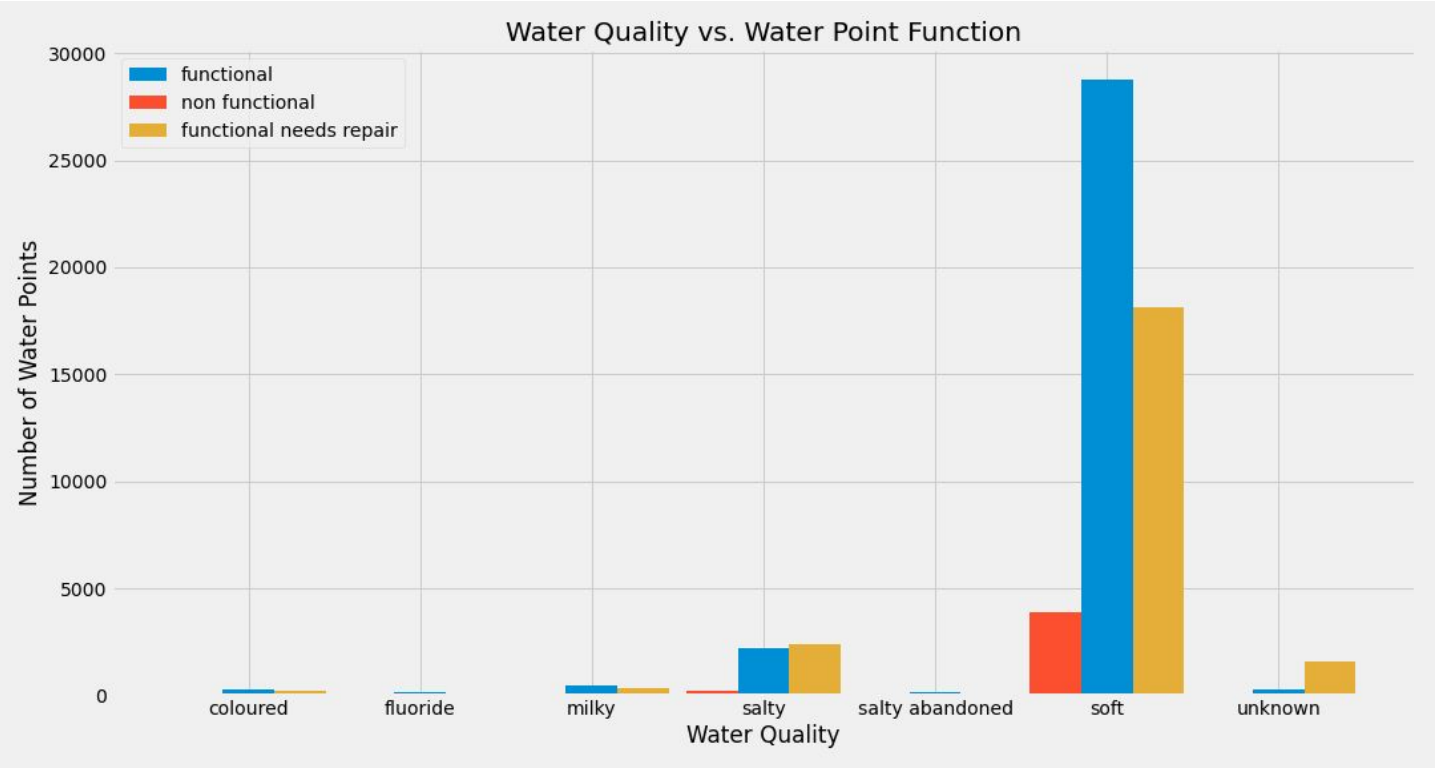
Do you have any questions?

Aaron Cherry
Email: cherrya050@gmail@.com
Github: https://github.com/JCherryA050
Linkedin: https://www.linkedin.com/in/aaron-cherry-8aa728124/

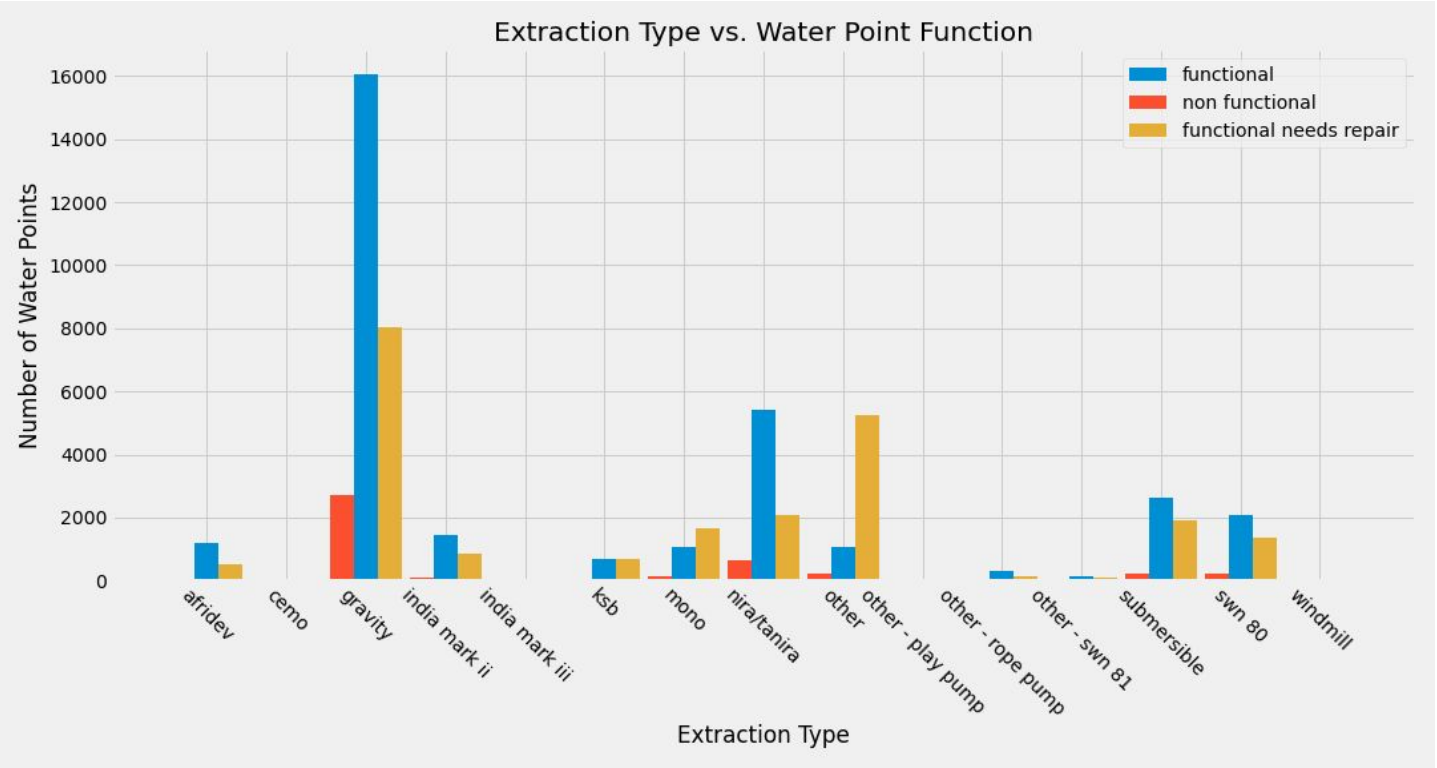
Victor Chen
Email: victor.i.chen.98@gmail.com
Github: https://github.com/vchen-98
Linkedin: https://www.linkedin.com/in/victorchen98/

Auxiliary Slides

Water Quality

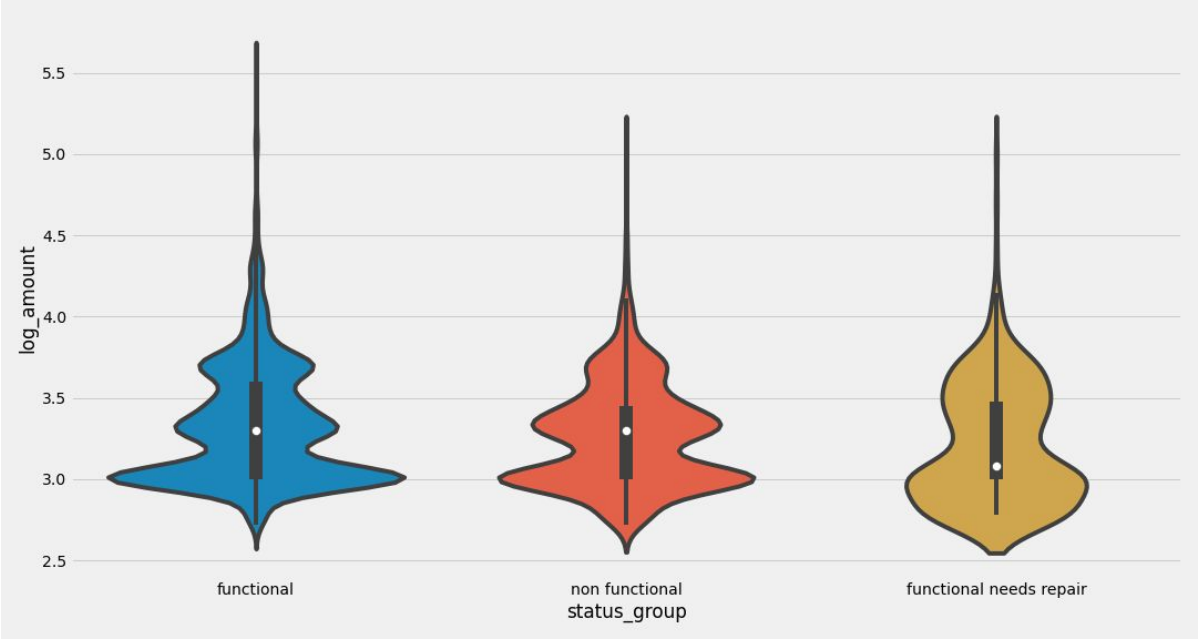


Operation Type

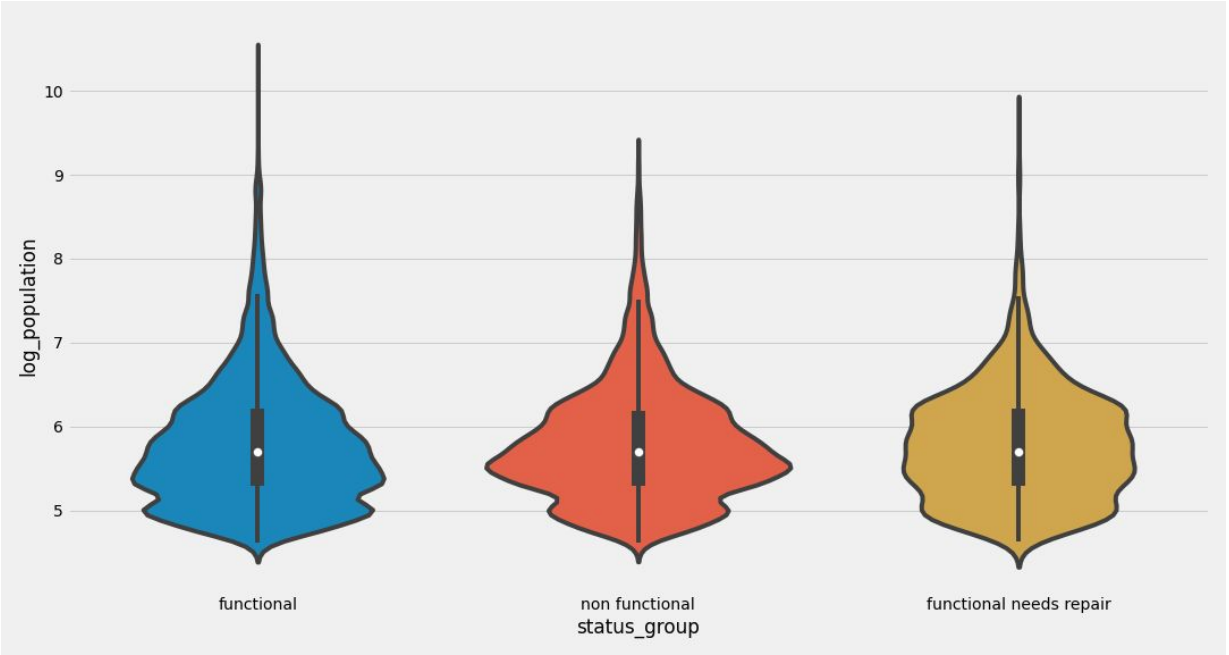


Water Yield

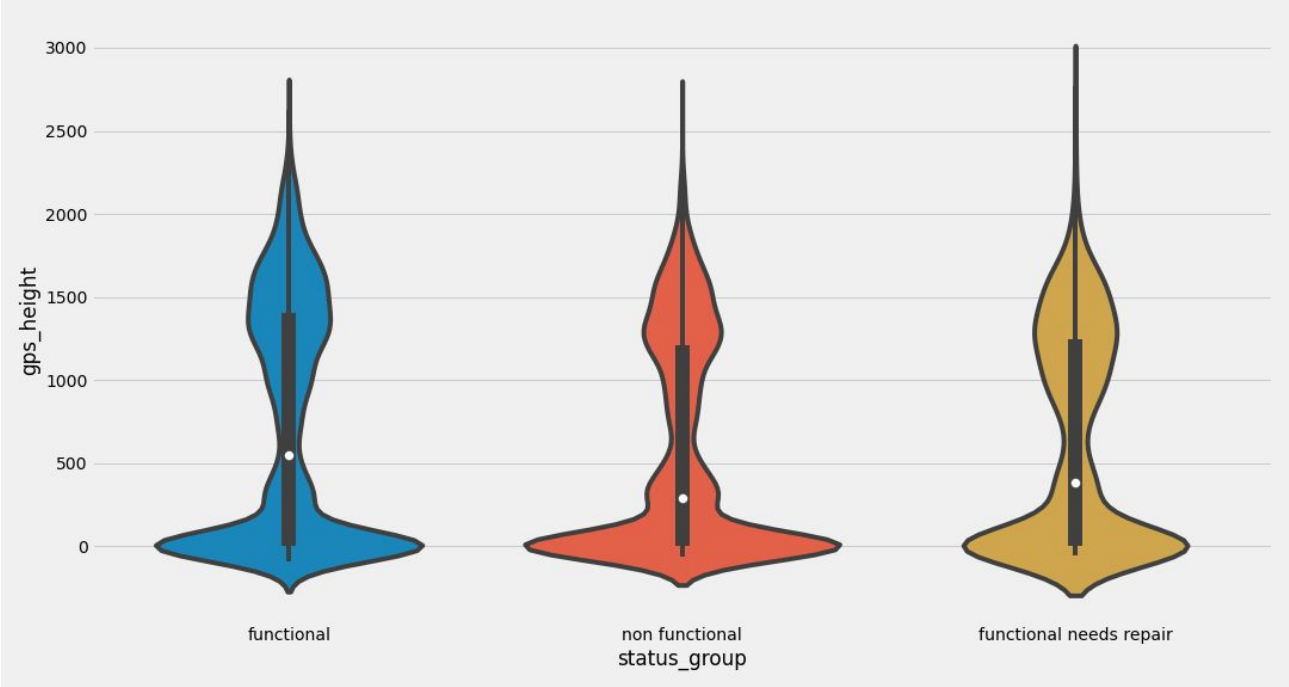
-


5.5
5.0
4.5
4.0
3.5
3.0
2.5
log_amount
functional
non functional
functional needs repair
status_group

Surrounding Population



Altitude


3000
2500
2000
1500
1000
500
0
gps_height
functional
non functional
functional needs repair
status_group