WATER PUMPS CLASSIFICATION

Abigail Campbell

PROMOTING ACCESS TO CLEAN WATER ACROSS TANZANIA

Can we predict which water pumps need maintenance?

Water pumps installed throughout the country	Maintaining these water pumps can prove to be a challenge	Many in remote areas that are difficult to monitor

Goals:

1. *Create a model that predicts the maintenance requirements of the water pumps*
    1. *Final model precision: 80.5%*
2. *Determine the most important features involved in predicting the functionality*
    1. *Location, quantity of water, construction year*

WHAT DATA WILL BE USED TO TRAIN THE MODEL?

Taarifa waterpoints dashboard

- aggregates data from the Tanzania Ministry of Water
- <53,000 pumps in the training data set (after cleaning)

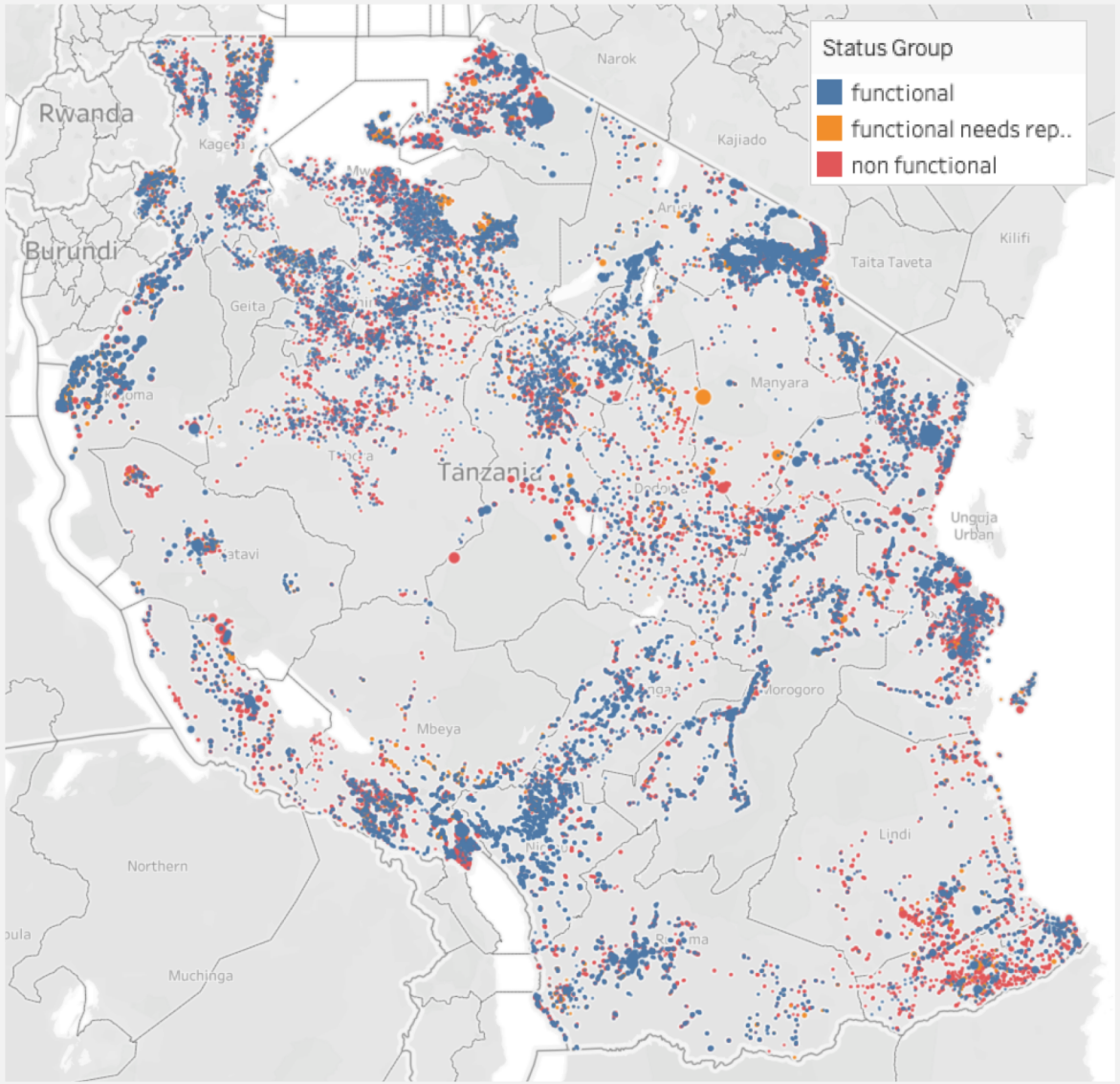
Location

Characteristics

People

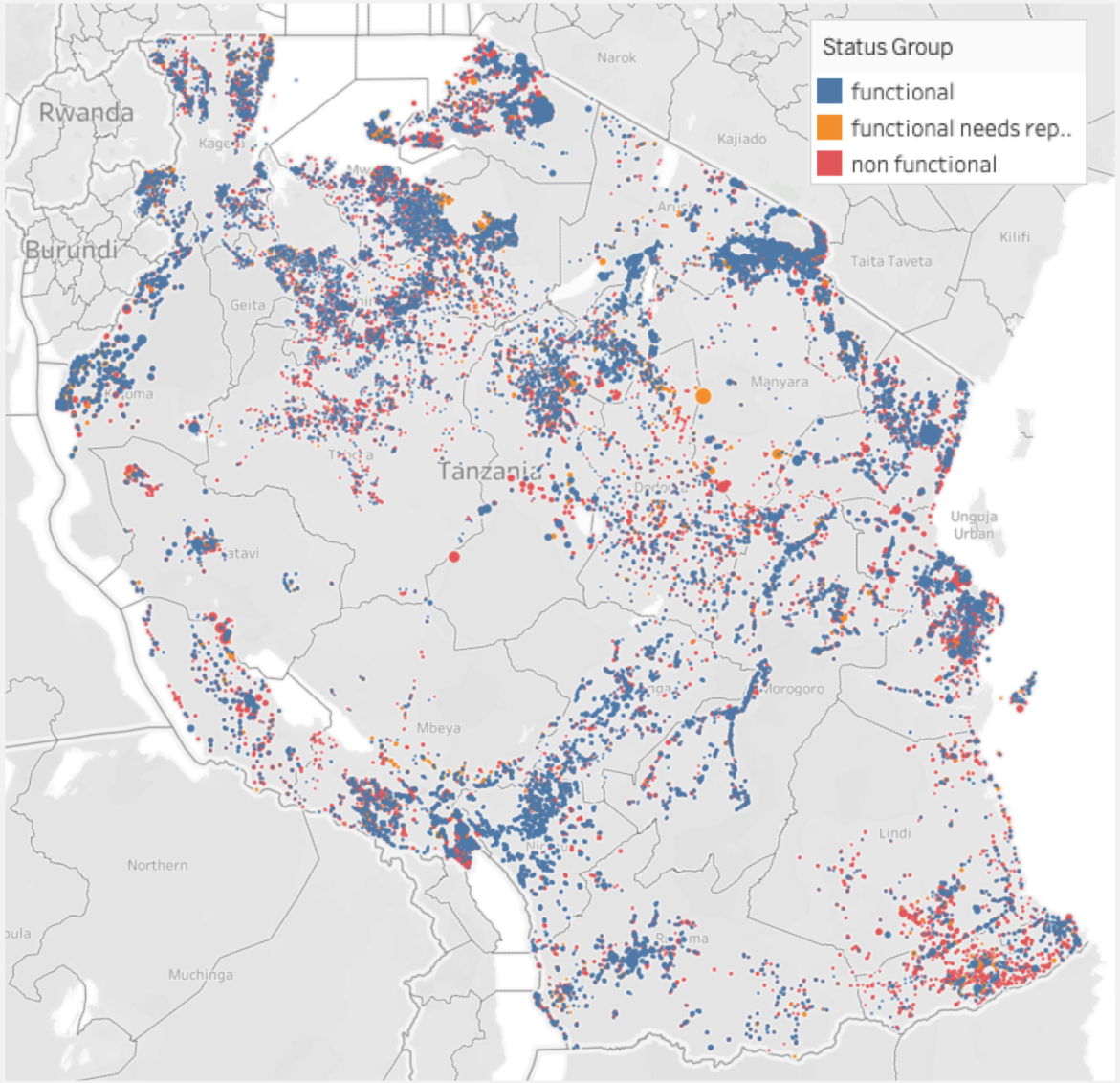
Water Source

A LOOK INSIDE THE DATA

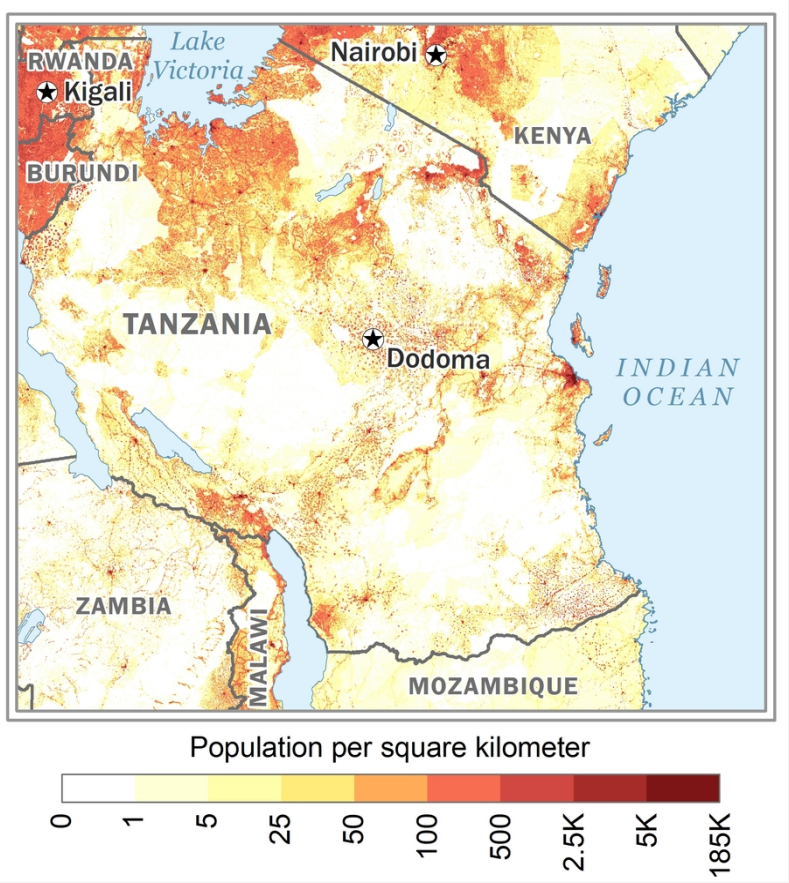


Locations of pumps across Tanzania

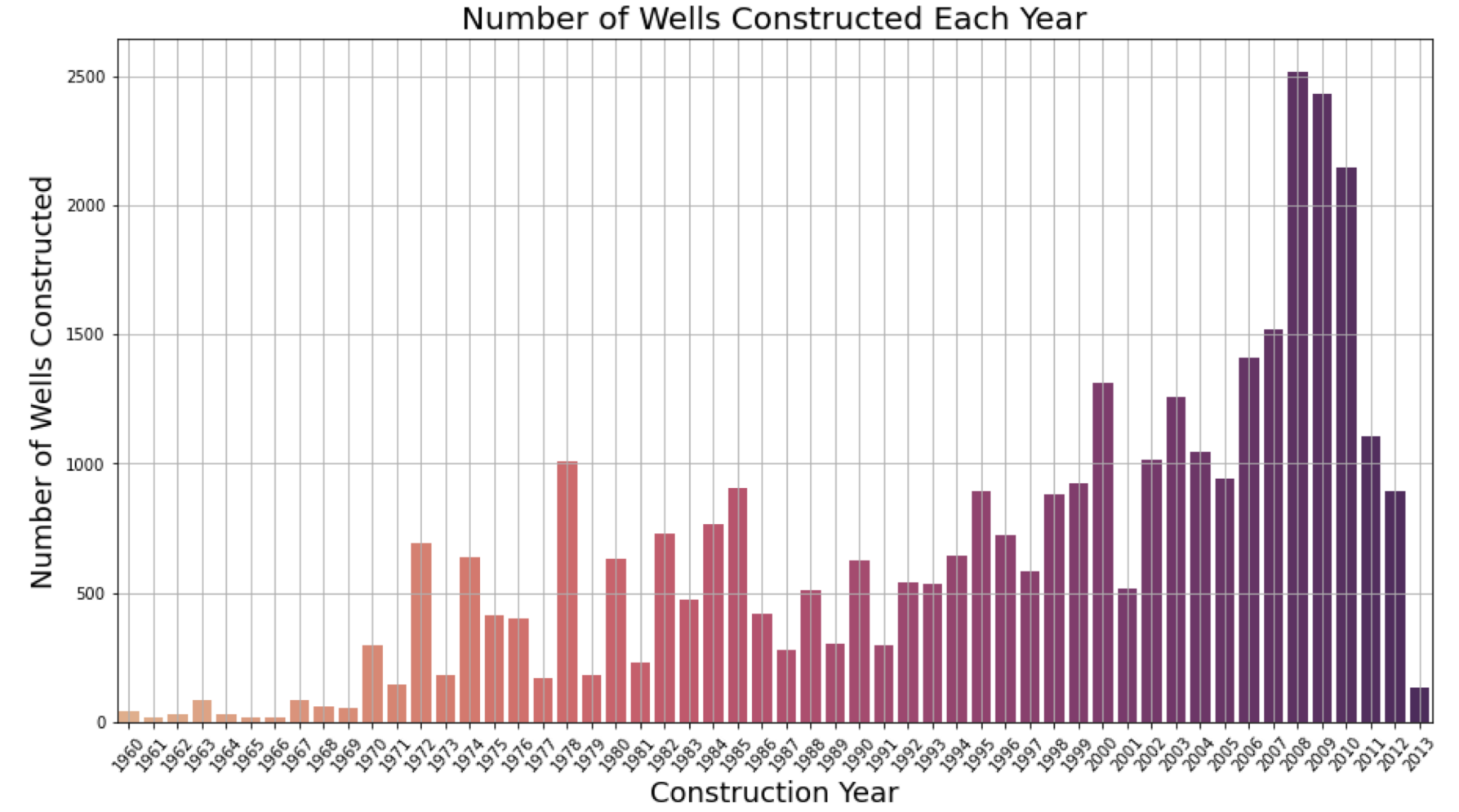
A LOOK INSIDE THE DATA



Locations of pumps across Tanzania



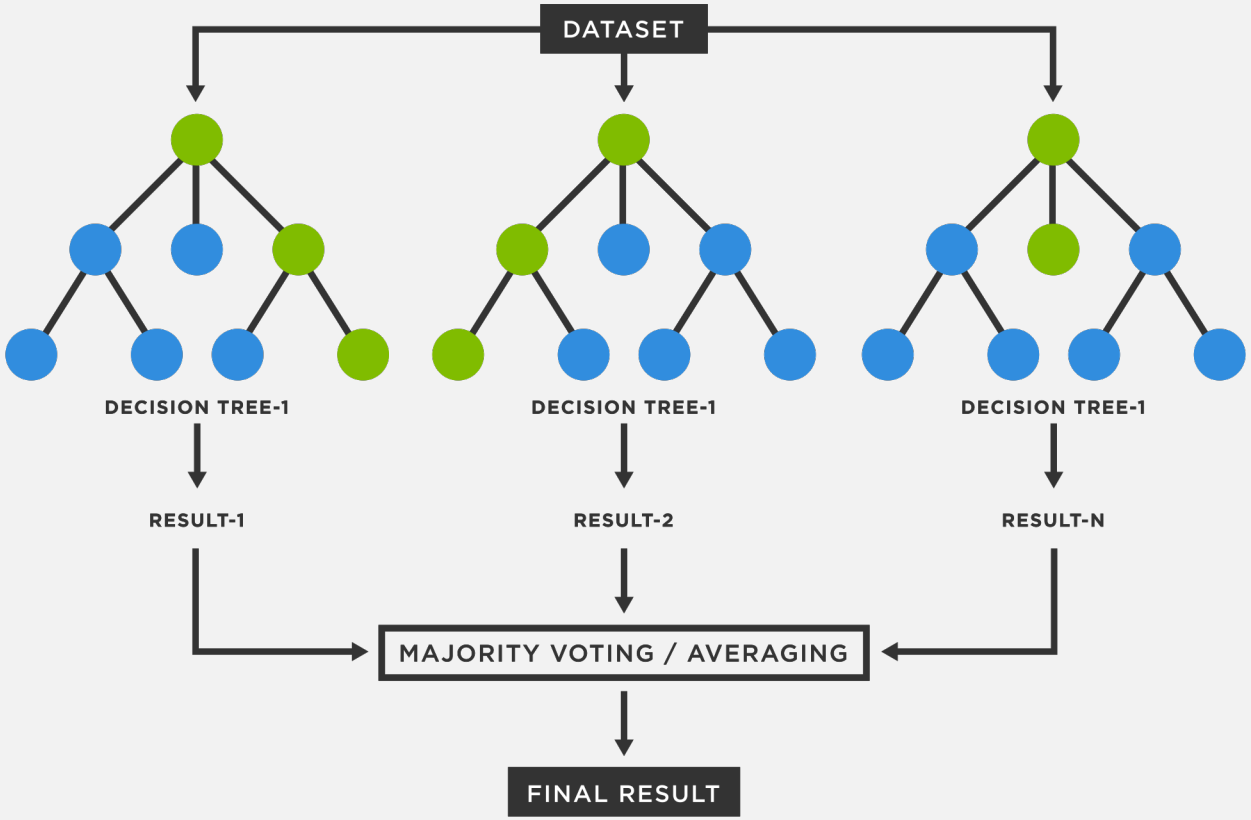
A LOOK INSIDE THE DATA



Steady increase in the number of pumps constructed each year

MODEL SELECTION

Model: Random Forest Classifier

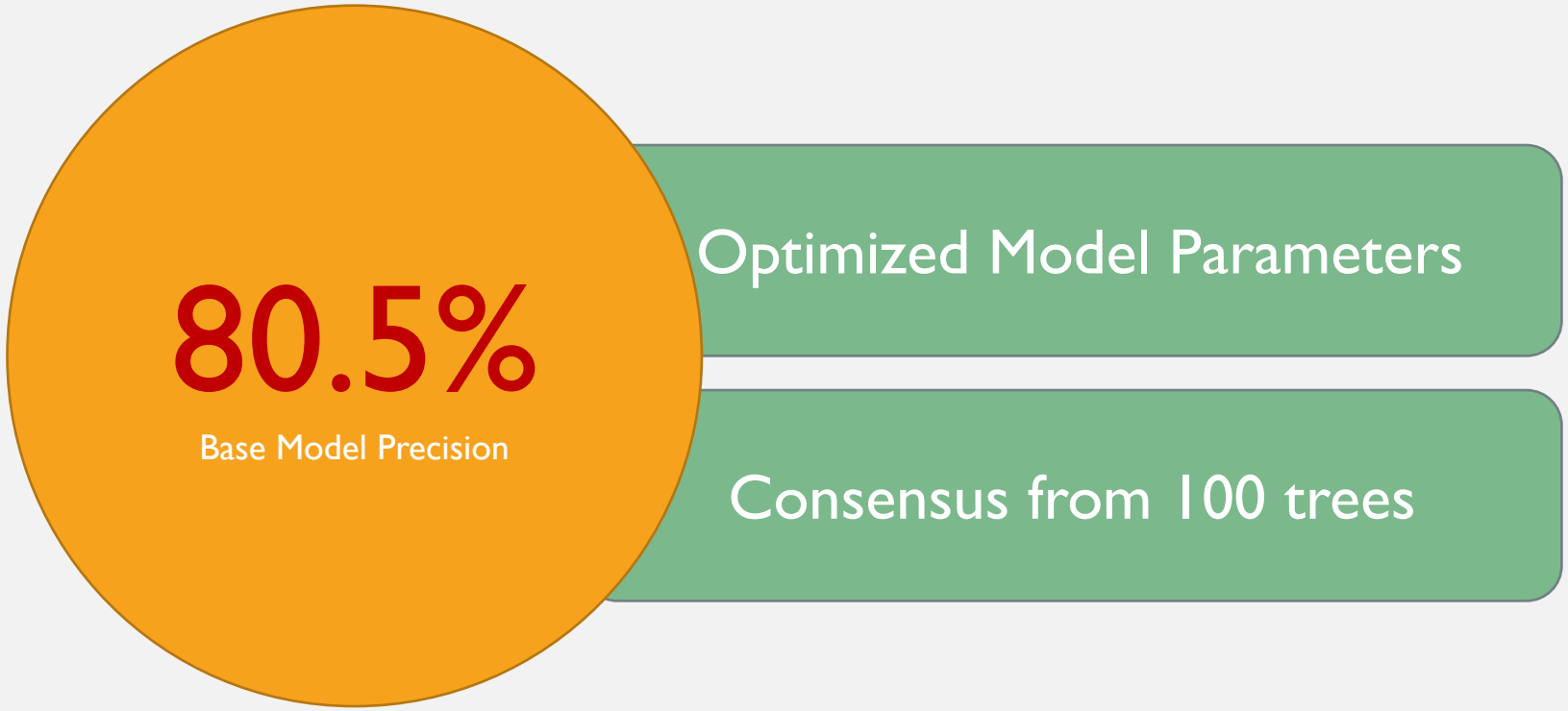


METRIC SELECTION

Metric: Precision

- Reducing false positives
- Better to visit a well that doesn't need fixing rather than miss a well that does

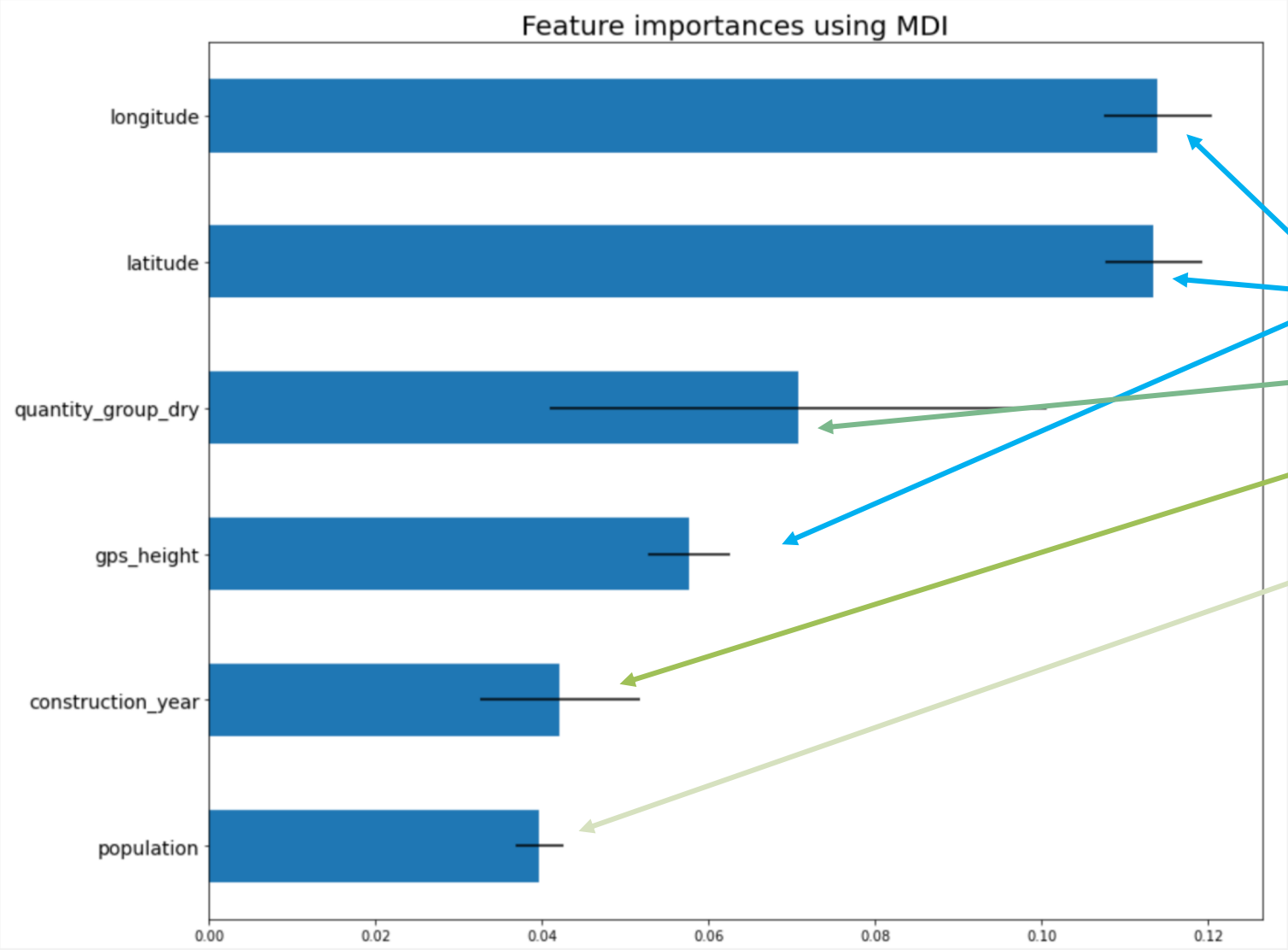
RANDOM FOREST MODEL PERFORMANCE



Additional Metrics:

Accuracy	Recall	F1 Score
82.1%	88.4%	84.3%

WHICH PUMP FEATURES ARE THE MOST IMPORTANT?



Most important Features

1. Location
2. Quantity of Water
3. Construction Year
4. Population Size

CONCLUSIONS

1. Our model can predict the functionality of a up with up to 80.5% precision

2. The most important factor in identifying functional pumps is location.
    1. Identify regions with high numbers of non-functional pumps to focus on first.

3. The next most important factors are water quantity and construction year.
    1. Prioritize old pumps and pumps with “dry” quantities

WHAT NOW?

Create a Priority Ranking

- Most likely to be non functional
- Serve the largest populations

Send out Crews

- Send out maintenance crews to the highest priority pumps

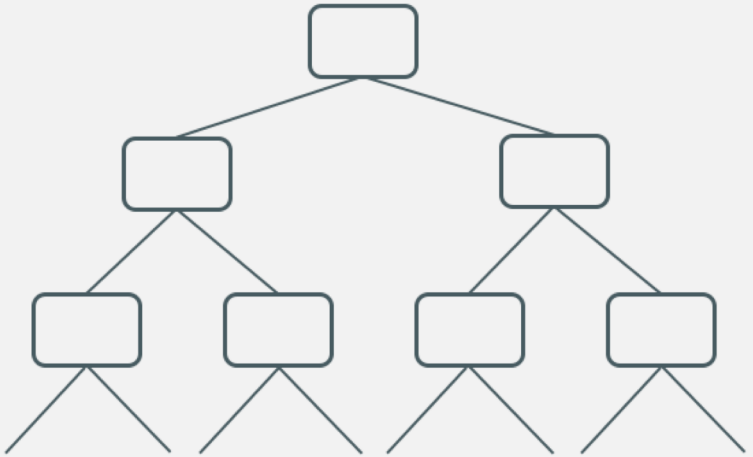
CONTACT INFORMATION

Abigail Campbell

abbycampbell0@gmail.com
(801) 541-2771

APPENDIX

MODEL AND METRICS SELECTION



Decision Tree Classifier

- ~~Numerical and Categorical data~~
- ~~Multiple possible labels~~
- Simple

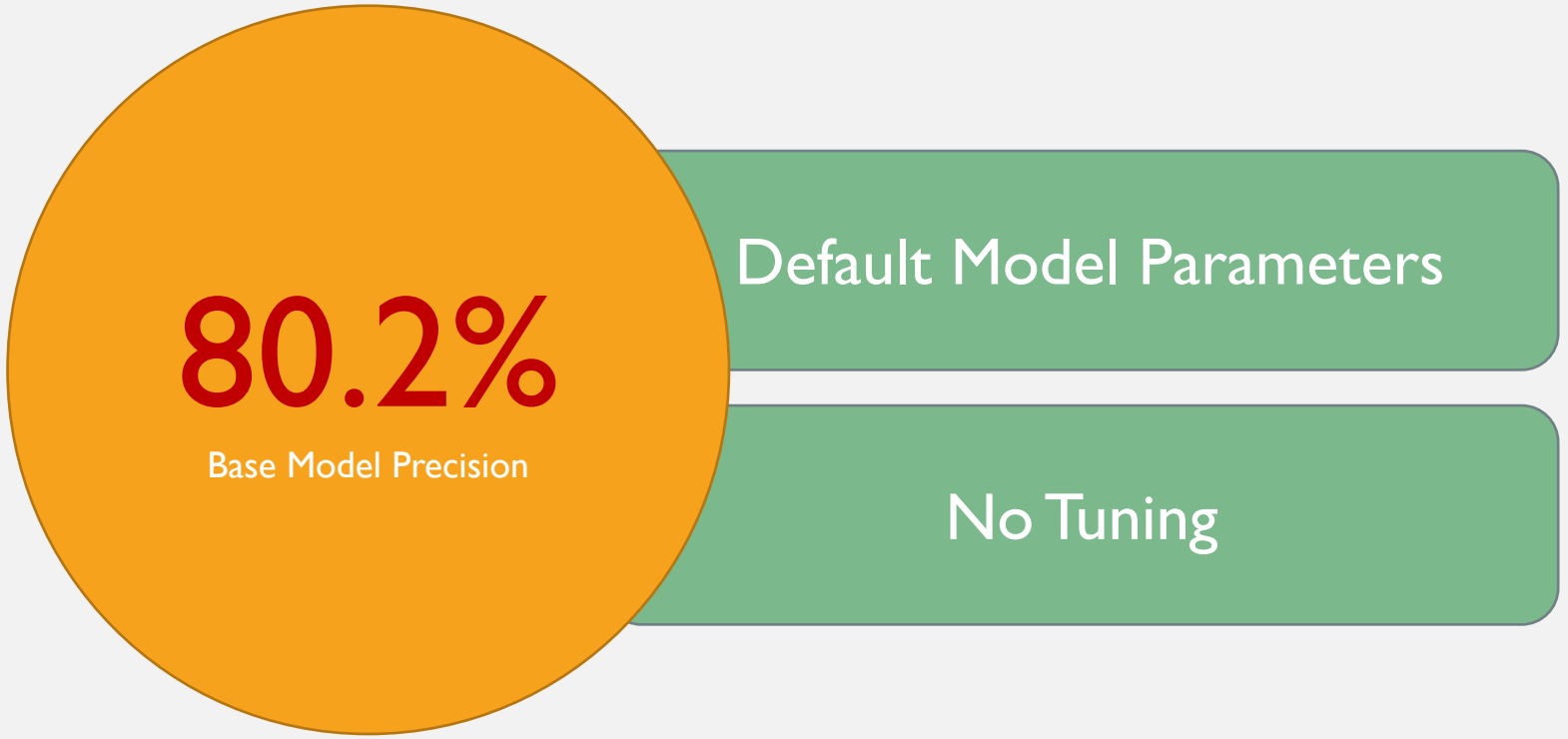
Criterion: Gini

- Measures frequency of mislabels
- Typically faster and less complex than other criterion

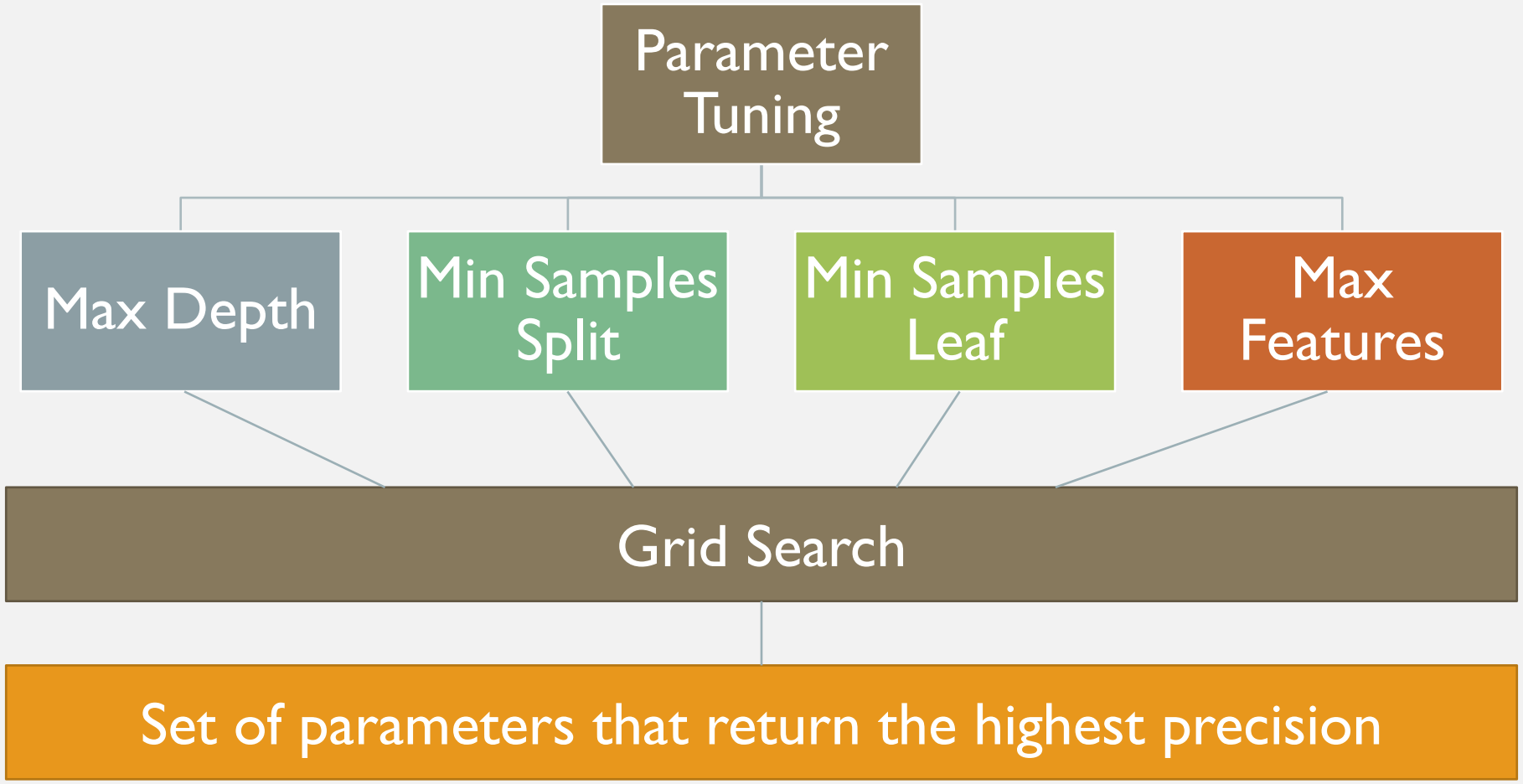
Metric: Precision

- Prioritizes reducing false positives
- Better to visit a well that doesn't need fixing rather than miss a well that does

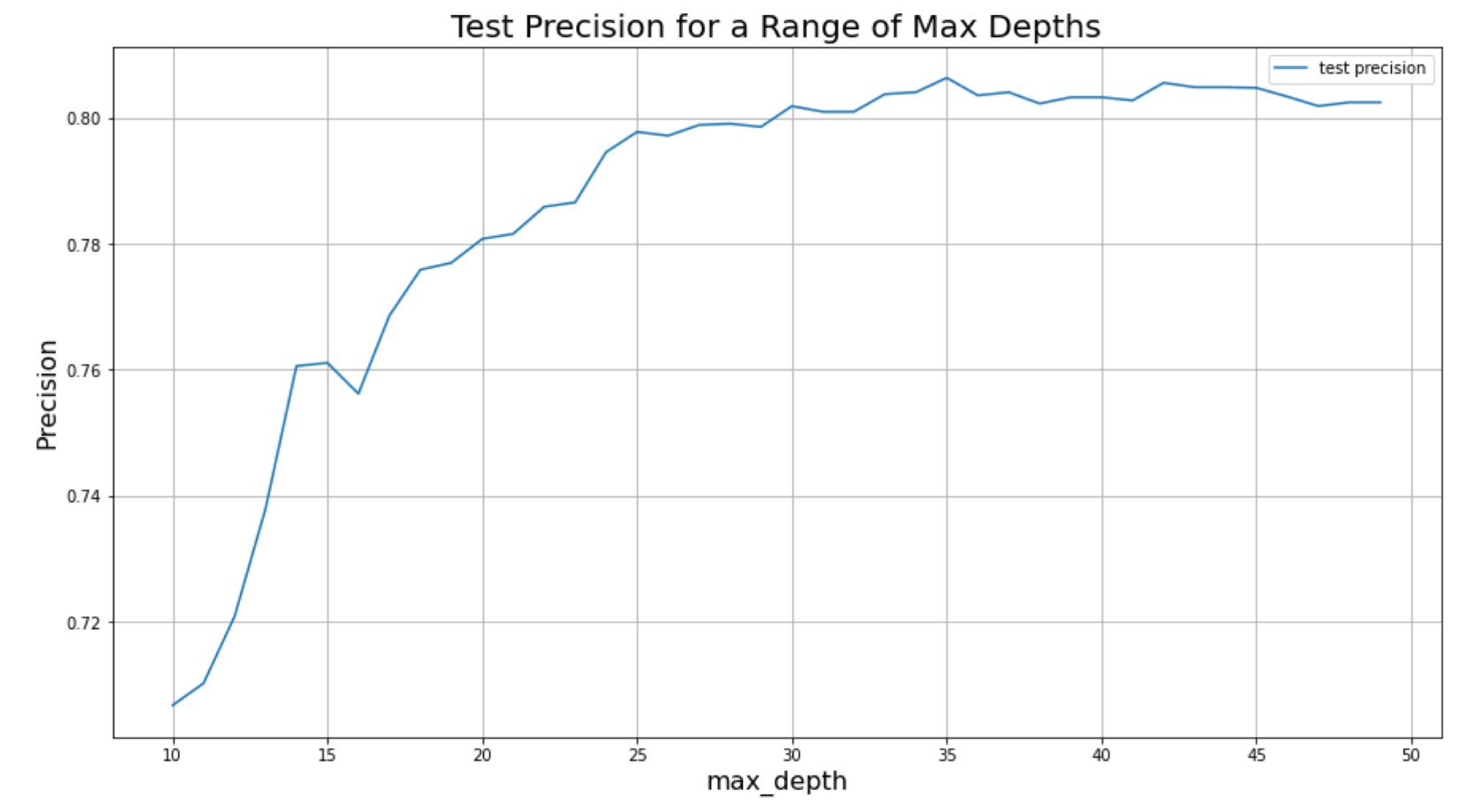
BASE MODEL


80.2%
Base Model Precision
Default Model Parameters
No Tuning

MODEL TRAINING



MODEL TRAINING – PARAMETER TUNING



The resulting precision from a range of parameter inputs is tracked, with the maximum precision value or range being selected.

MODEL TRAINING – GRID SEARCH

1. Every possible combination of parameters is generated
    2. Each combination is used to train a decision tree
        3. The resulting precision is tracked
4. The combination resulting in the highest precision is selected

Parameter 1 (1-4)

Parameter 2 (A-D)				
	A1	A2	A3	A4
	B1	B2	B3	B4
	C1	C2	C3	C4
	D1	D2	D3	D4

2 parameter example

FINAL OPTIMIZED MODEL PARAMETERS

- 39 Max Depth
- 2 Min Samples Split
- 1 Min Samples Leaf
- 94 Max Features

- 47,520 Permutations tested
- Optimized for precision in identifying functional water pumps