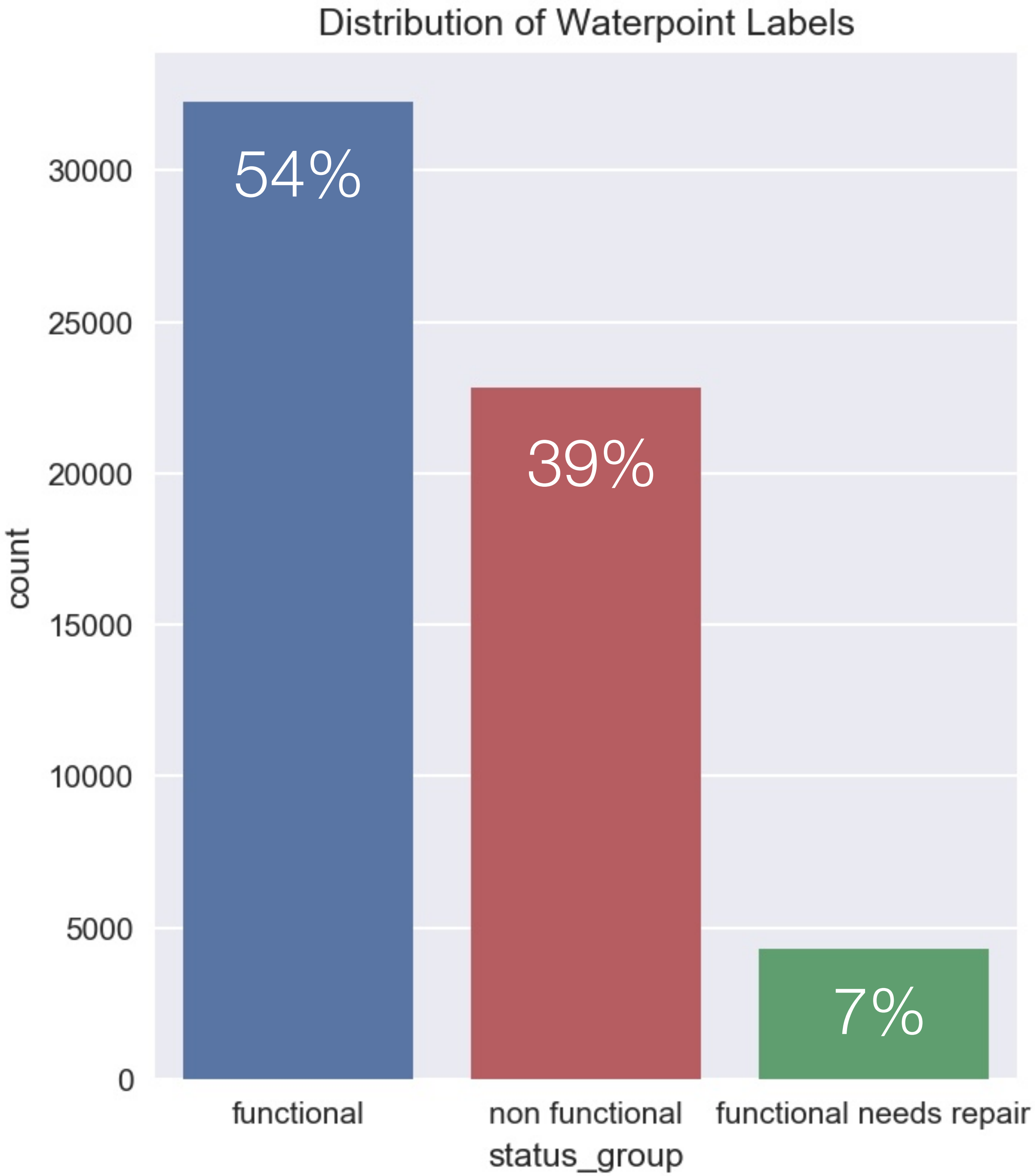
Pump It Up: Data Mining the Water Table

Predicting the operating condition of waterpoints in Tanzania

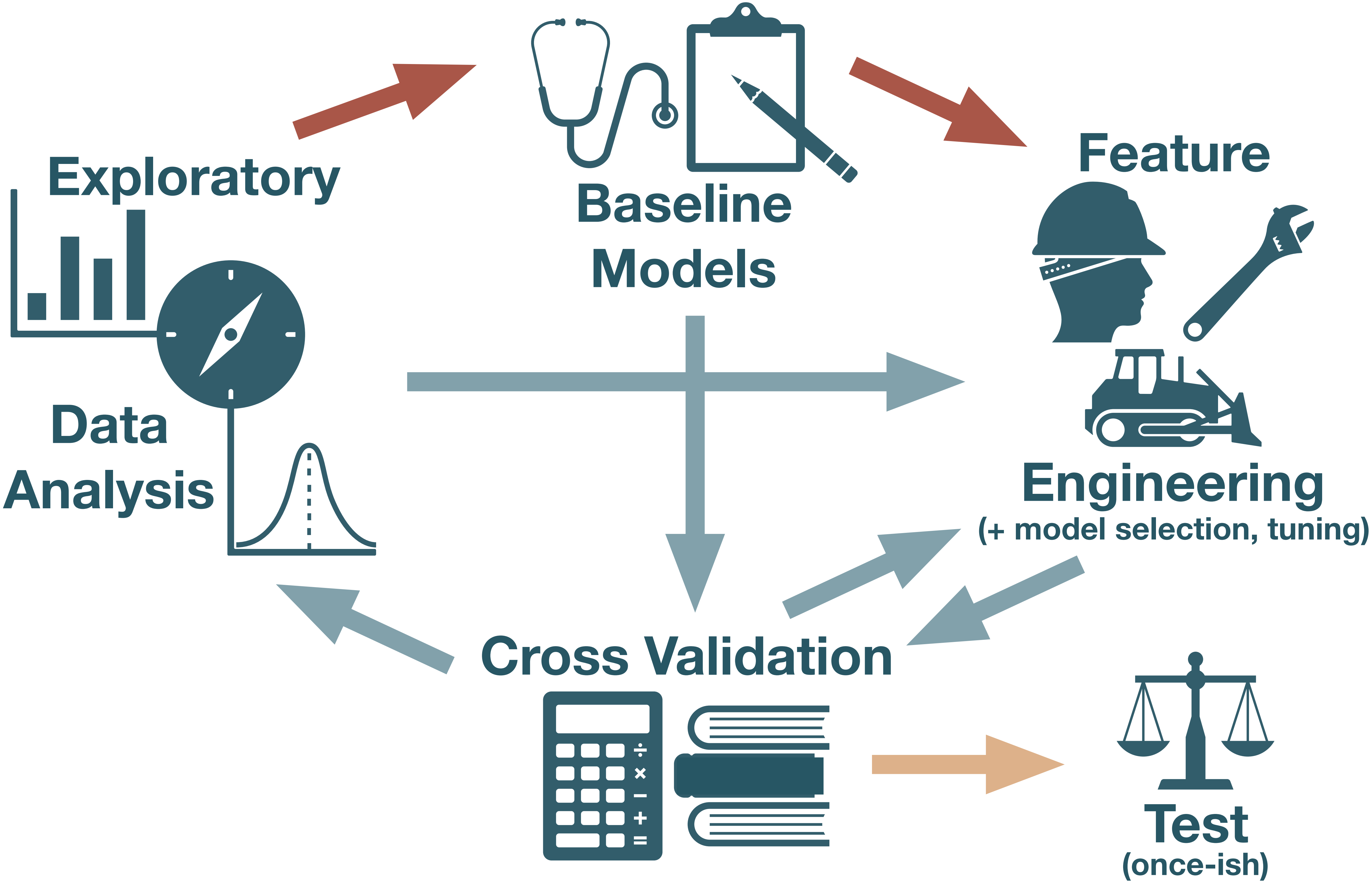
DRIVENDATA

Project Objective

- Multi-class problem using data from Taarifa and Tanzania Ministry of Water
- Target waterpoint labels: **functional**, **non functional**, & **functional needs repair**
- Training set of 40 features for 59,400 samples
- Test set of 14,850 unlabeled records
- Evaluation metric = classification rate (accuracy)



Classification Workflow



Exploratory Data Analysis

Baseline Models

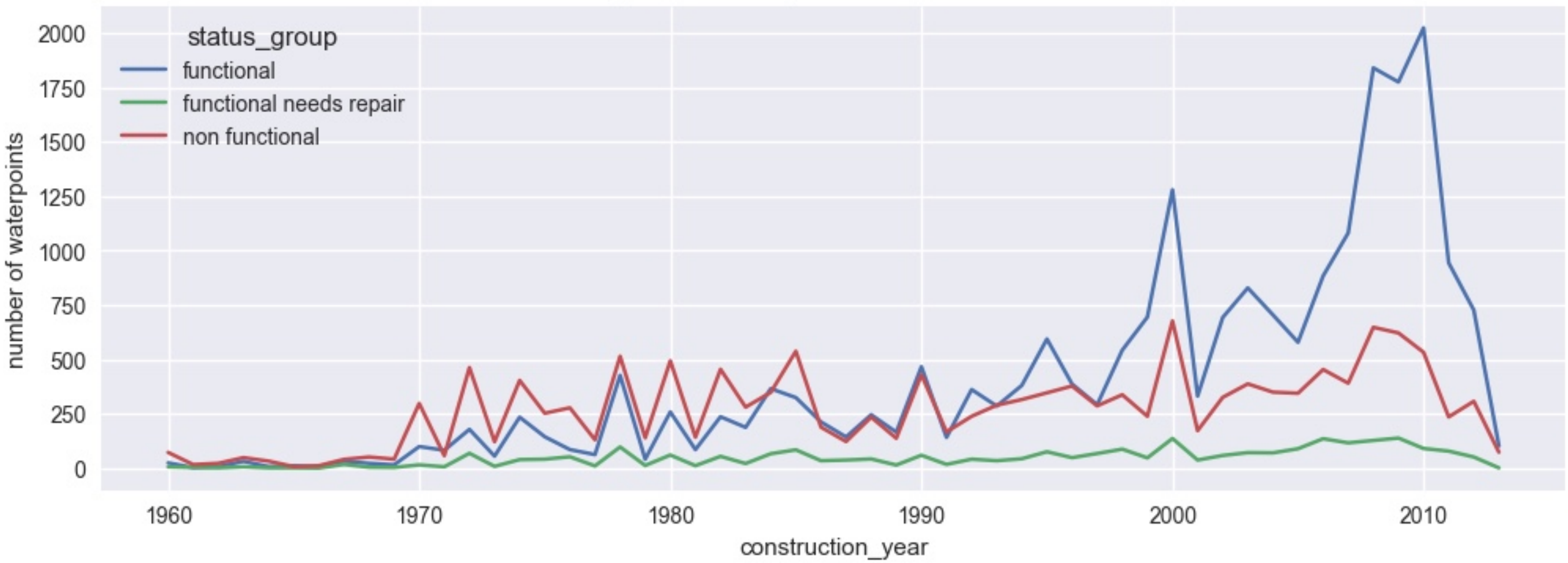
Feature Engineering
(+ model selection, tuning)

Cross Validation

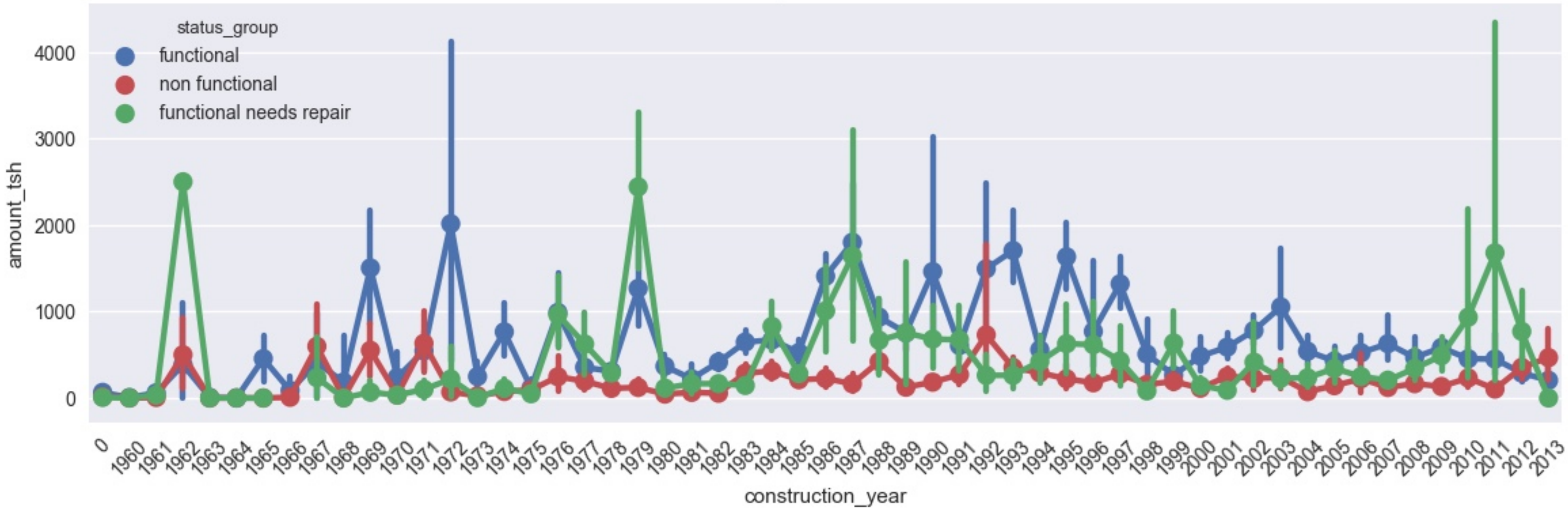
Test
(once-ish)

Exploratory Data Analysis - Continuous Features

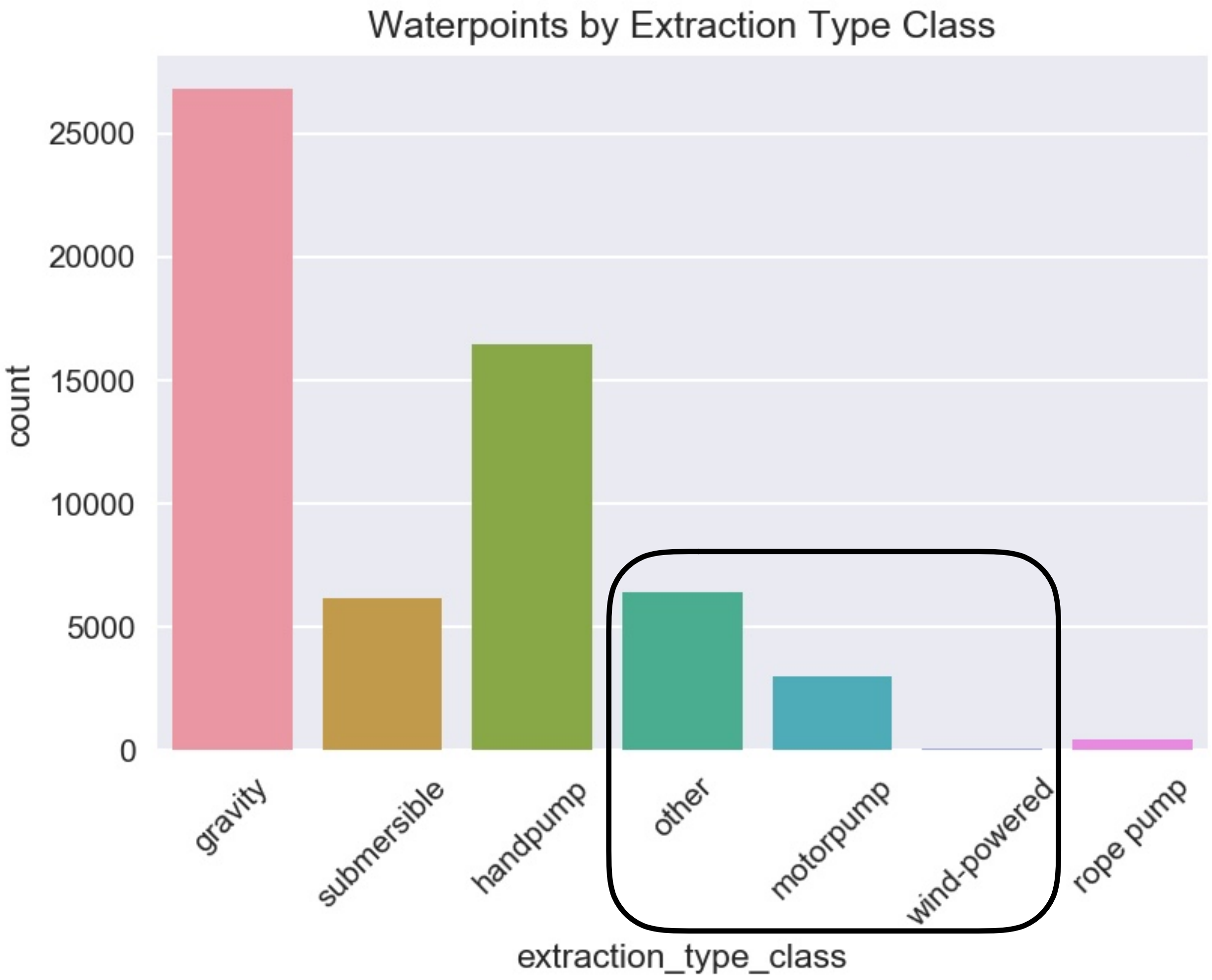
Waterpoint Counts by Status & Construction Year

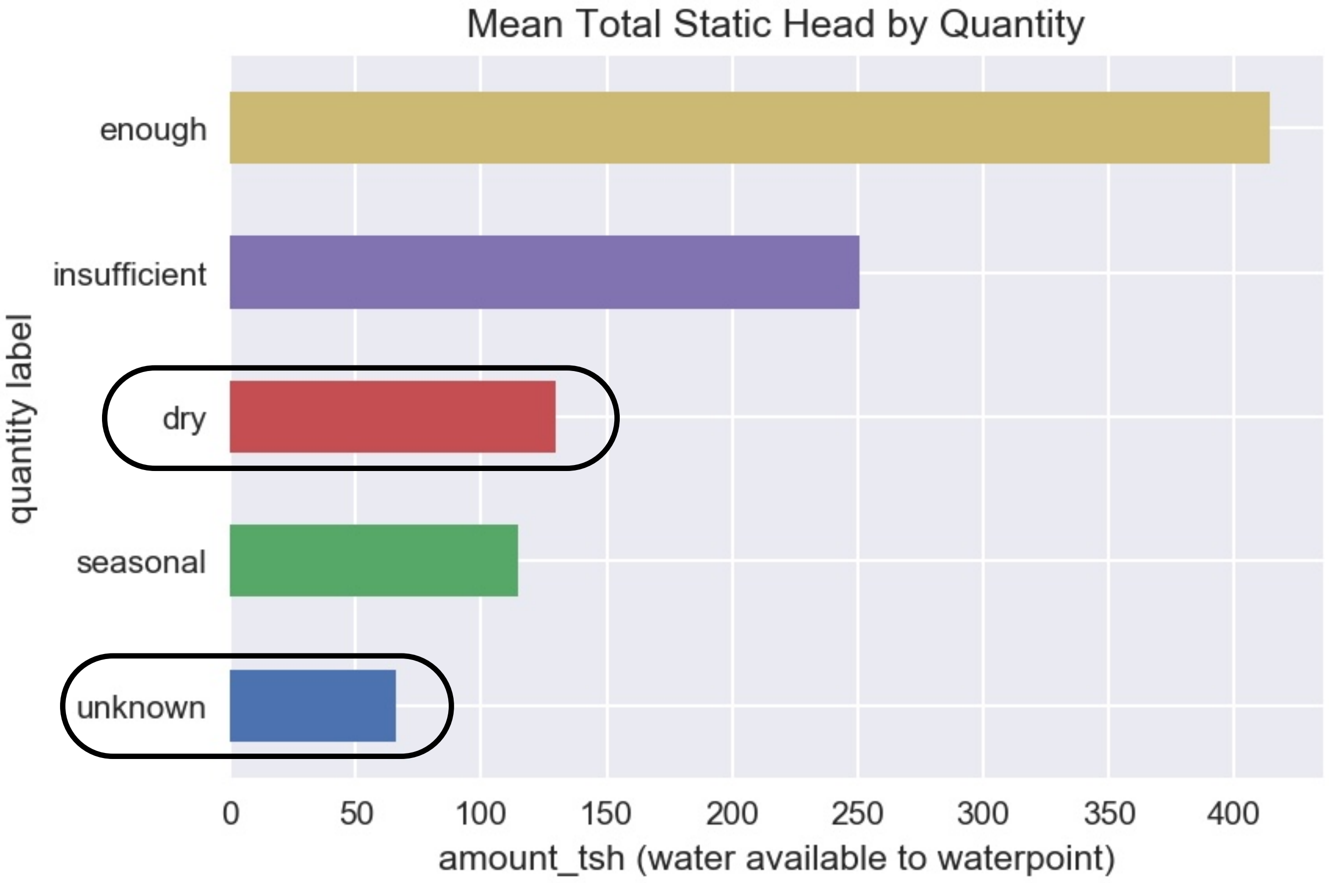


Mean Static Head by Status & Construction Year

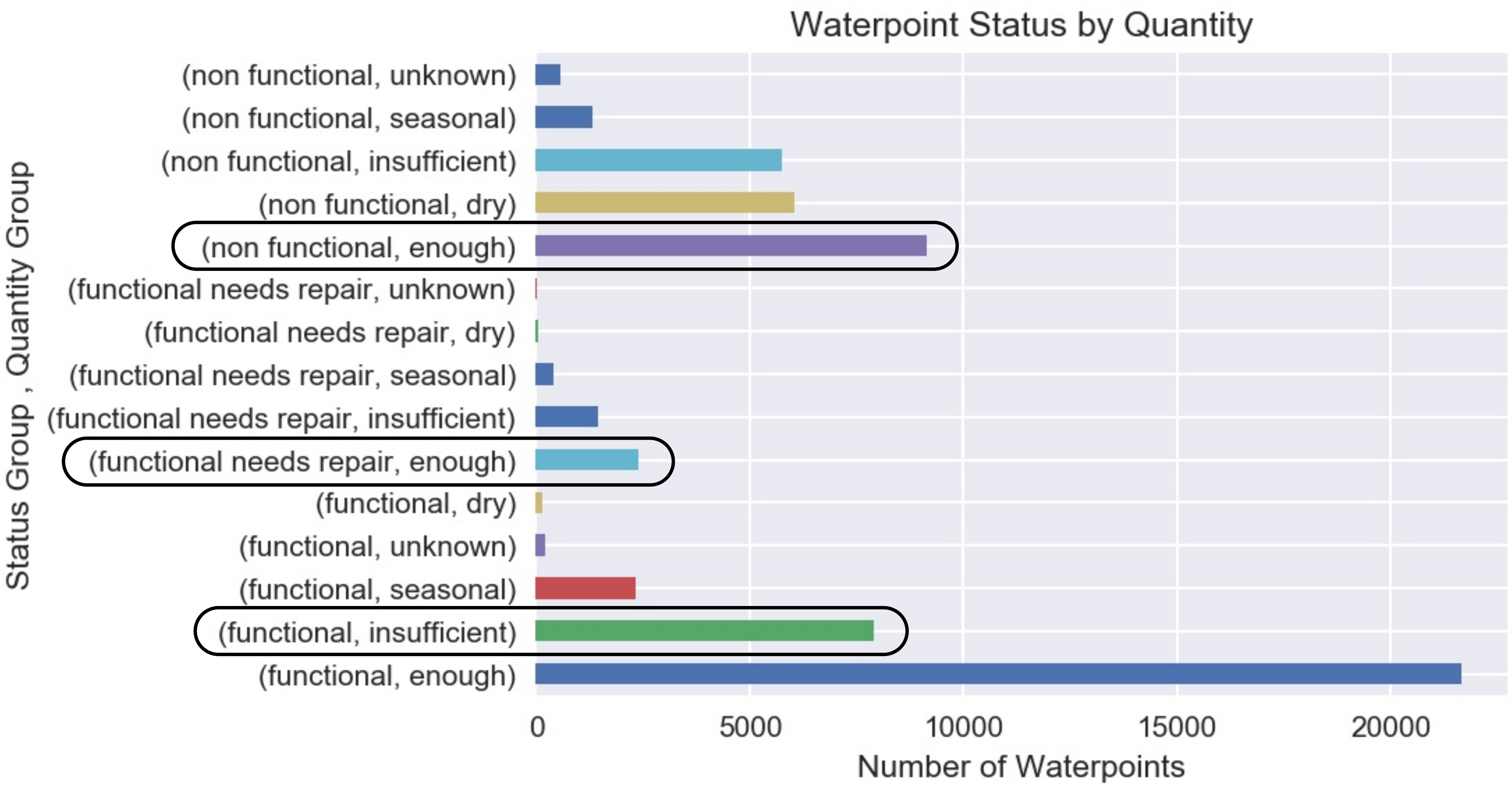


Exploratory Data Analysis - Categorical Features



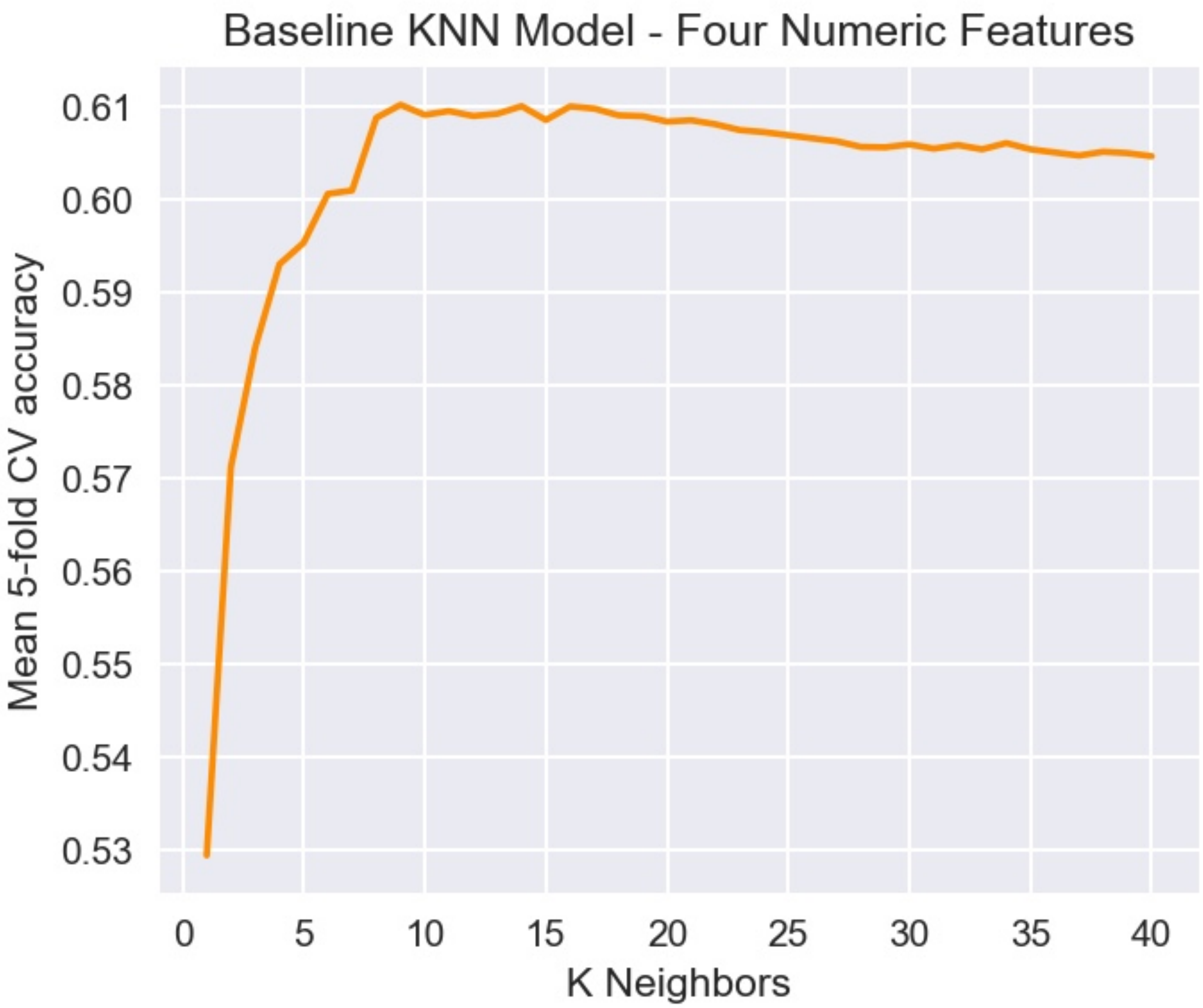


Exploratory Data Analysis - Categorical Features



Baseline Models

- Logistic Regression
  - 1 feature (tsh) = 0.542
  - 4 features (+ elevation, population, construction year) = 0.539
- K Nearest Neighbors
  - 4 features = 0.610
- Random Forest
  - 4 features = 0.642



Feature Engineering

12 of 30 categorical features ➔ over 4,000 dummies
…Good or bad idea?

Feature Engineering

12 of 30 categorical features ➔ over 4,000 dummies …Good or bad idea?

Representing latitude / longitude as 3 dimensions: x, y, and z ➔ Nearby values should be close in reality

Feature Engineering

12 of 30 categorical features ➔ over 4,000 dummies …Good or bad idea?

Representing latitude / longitude as 3 dimensions: x, y, and z ➔ Nearby values should be close in reality

2-feature combinations and ranges: tsh, quantity, population

Feature Engineering

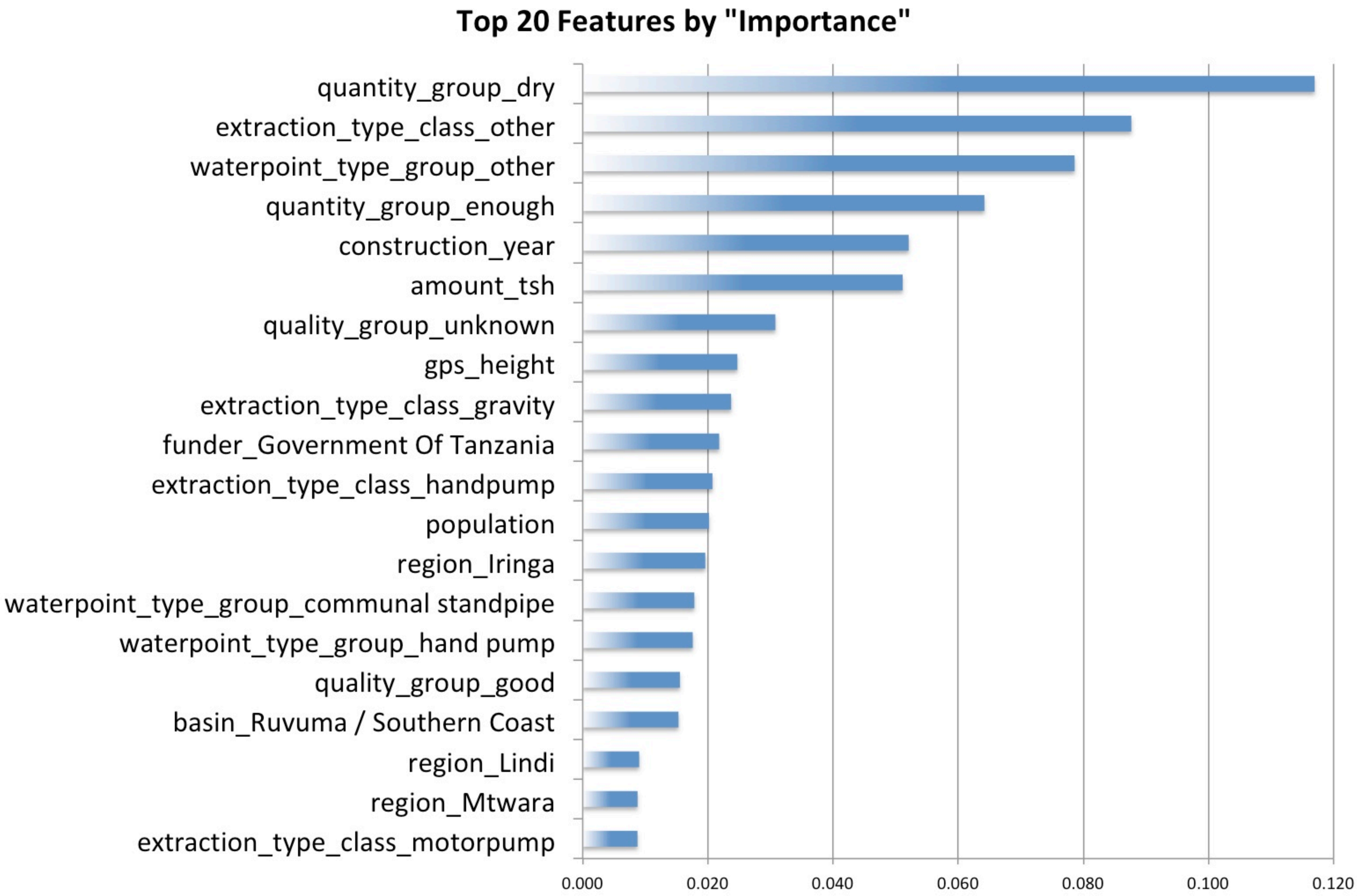
12 of 30 categorical features ➔ over 4,000 dummies …Good or bad idea?

Representing latitude / longitude as 3 dimensions: x, y, and z ➔ Nearby values should be close in reality

2-feature combinations and ranges: tsh, quantity, population

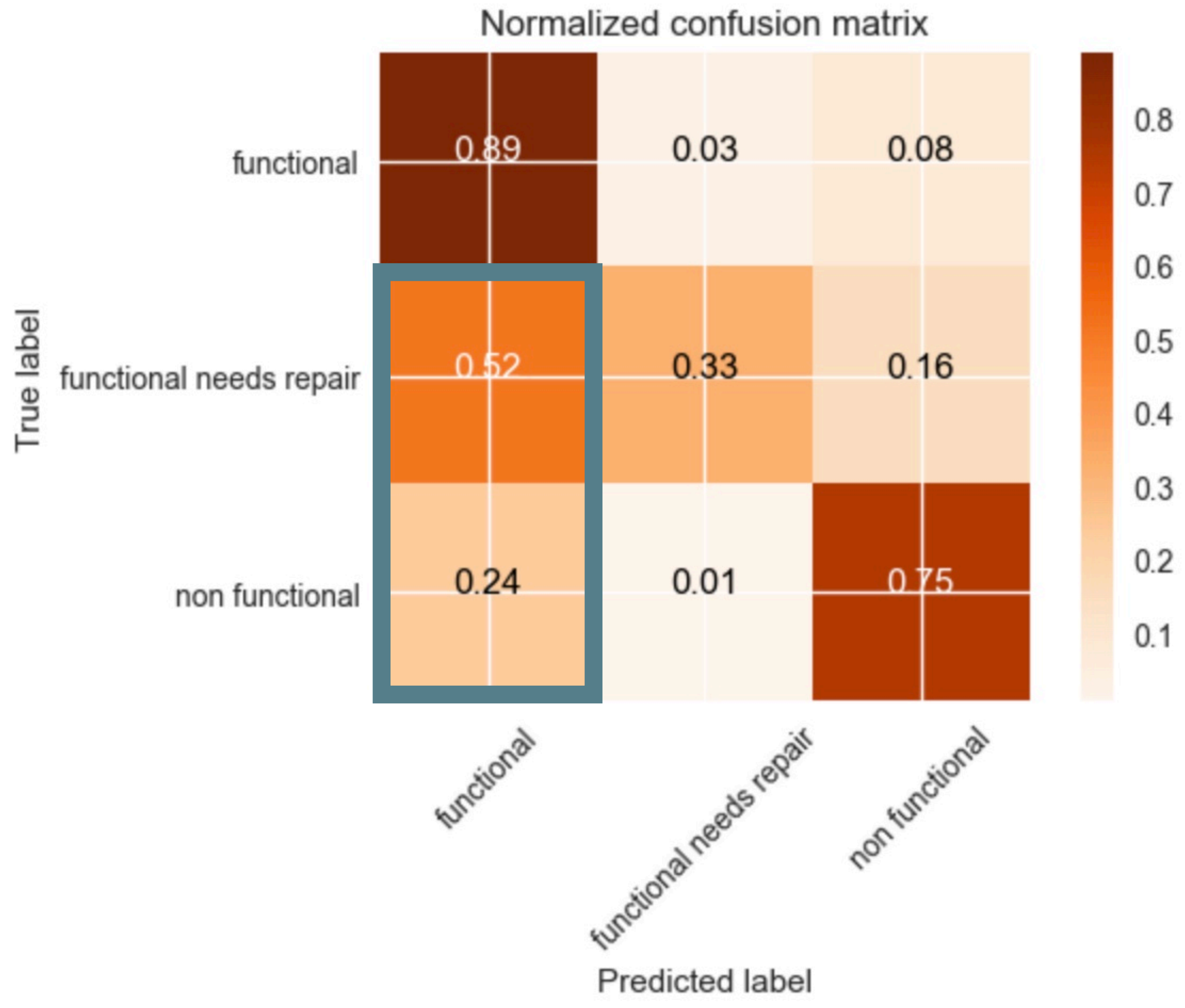
Does class imbalance matter? Not for accuracy, but for usefulness of model

Random Forest - Feature Selection & Model Tuning



CV Accuracy of 0.697 (4,000 features) to 0.736 (top 20 only)

Evaluating Final Model - What Does It Mean?



	functional	functional needs repair	non functional
functional	0.89	0.03	0.08
functional needs repair	0.52	0.33	0.16
non functional	0.24	0.01	0.75

99 features, 7 continuous
Accuracy = 0.802
F1 = 0.795 (weighted)
0.685 (macro)

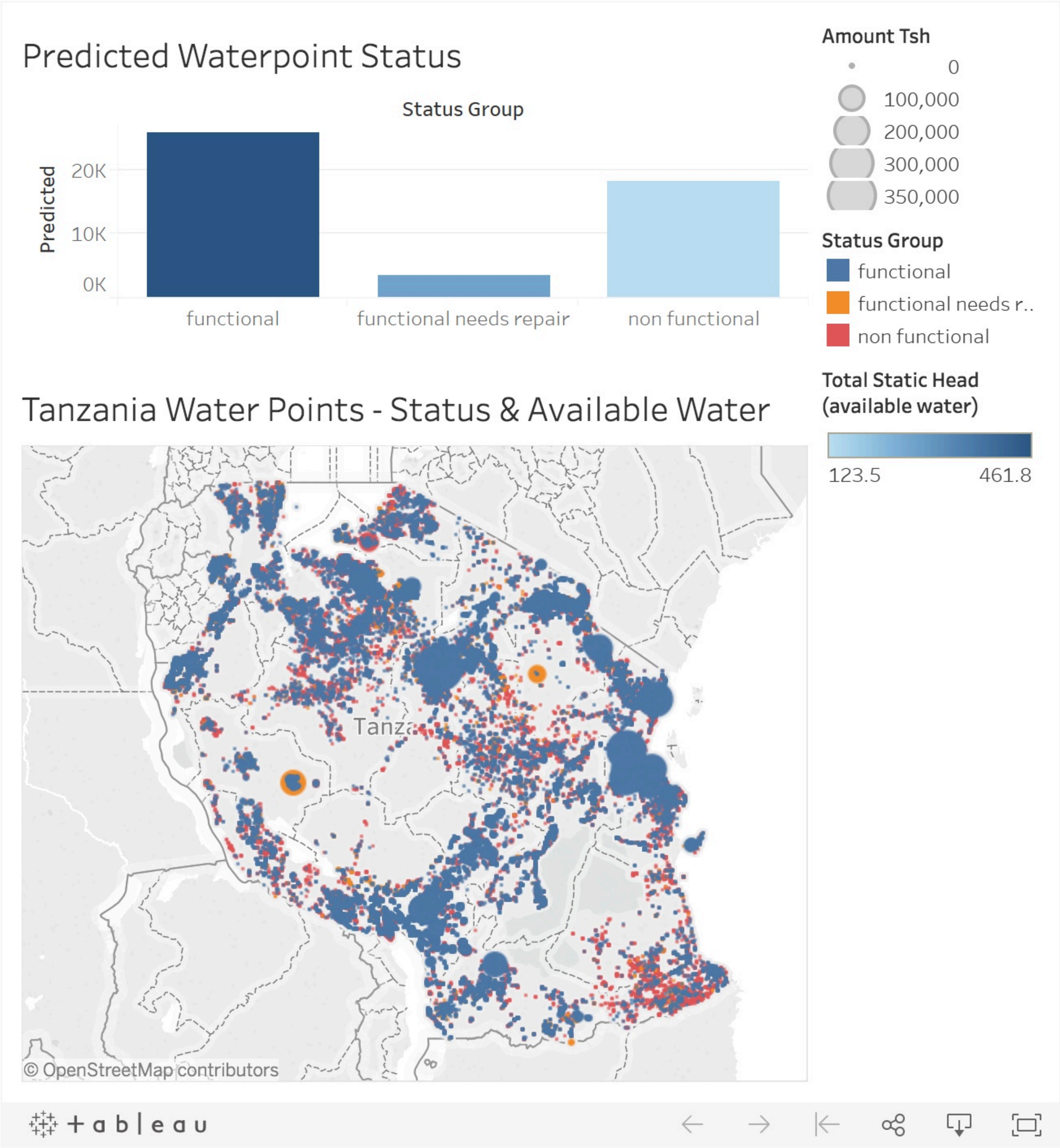
Factors to Consider

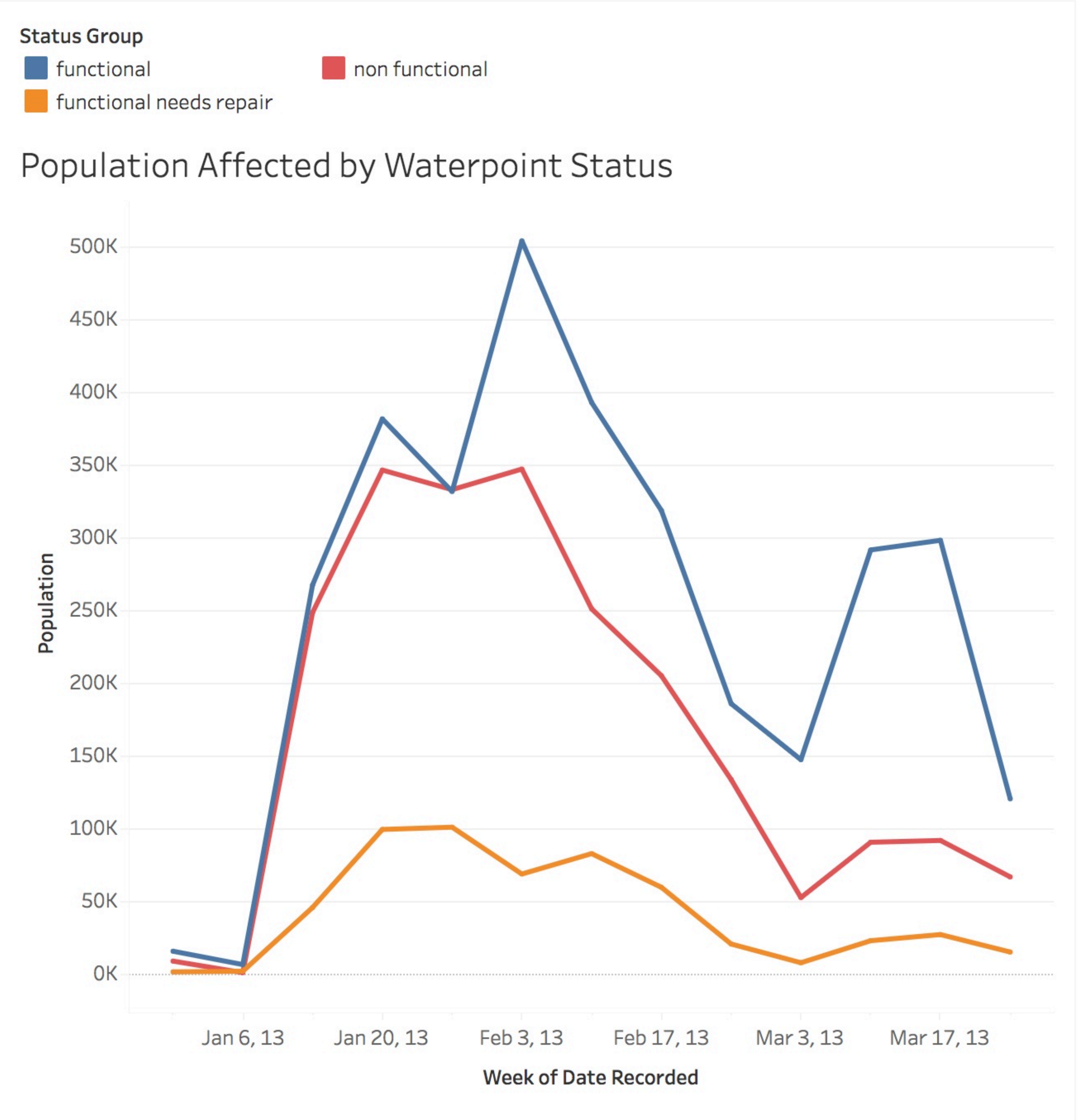
- Cost of Installation: $20-35,000
- Operation & Maintenance: ?
- Reliability
- Surrounding Population

Non Functional FN = $2,779,920
(assume $2k service)

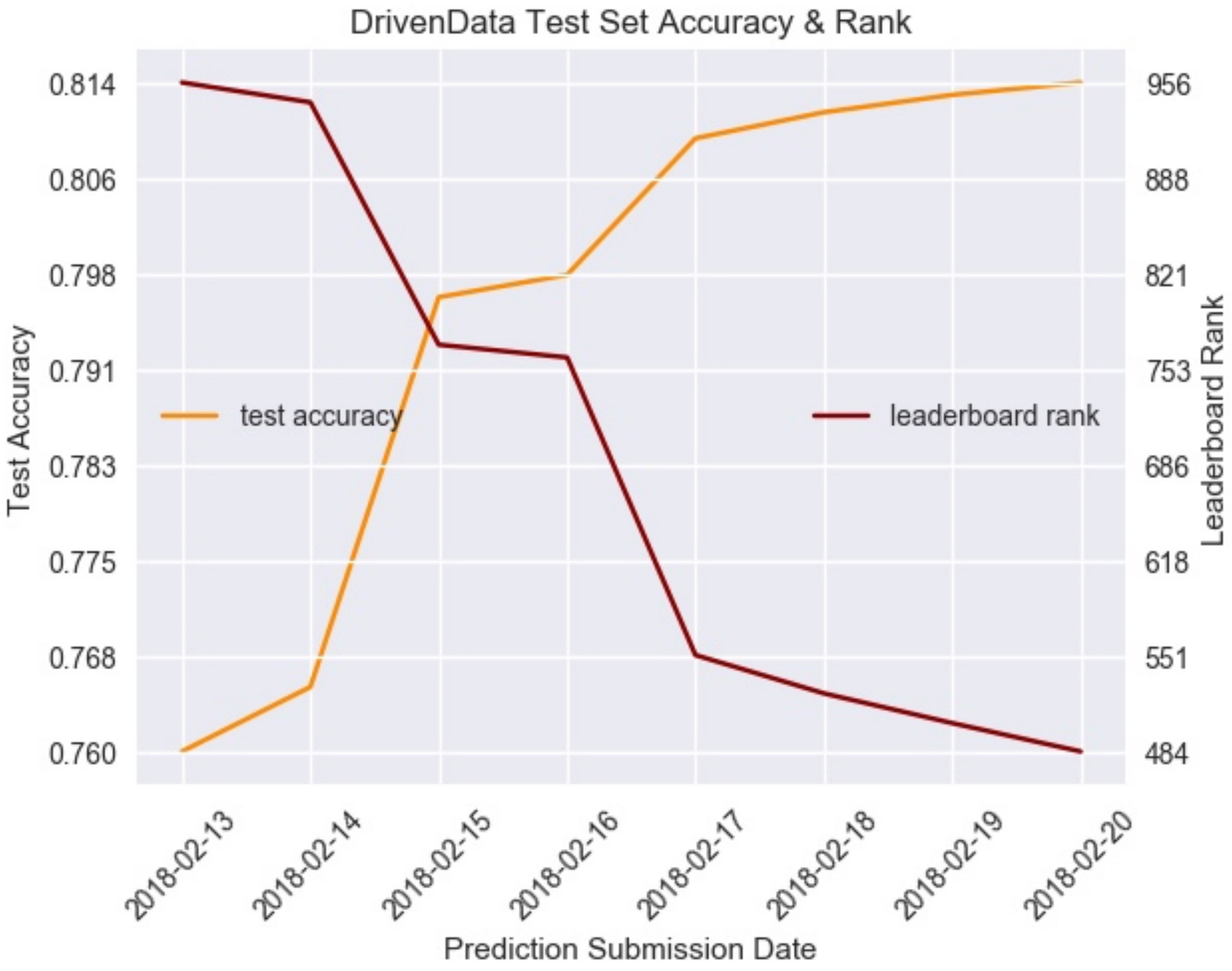
Needs Repair FN = $540,540
(assume $1k service)

Implementing Model - How Can We Use It?





DrivenData Submission Results


DrivenData Test Set Accuracy & Rank
Test Accuracy
Leaderboard Rank
test accuracy
leaderboard rank
Prediction Submission Date