Tanzania Water Well Classification

Multilabel Classification

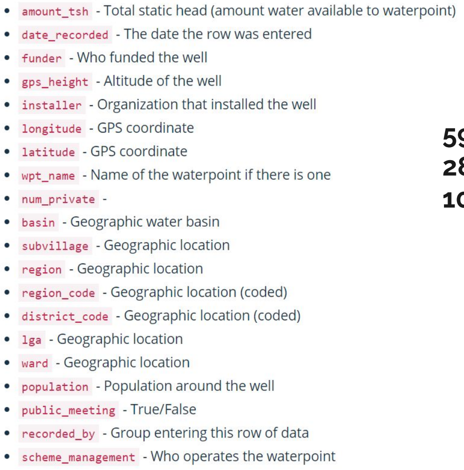
Background

- The data for this competition comes from the Taarifa waterpoints dashboard, which aggregates data from the Tanzania Ministry of Water.

- Using data from Taarifa and the Tanzanian Ministry of Water, can you predict which pumps are functional, which need some repairs, and which don't work at all?A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

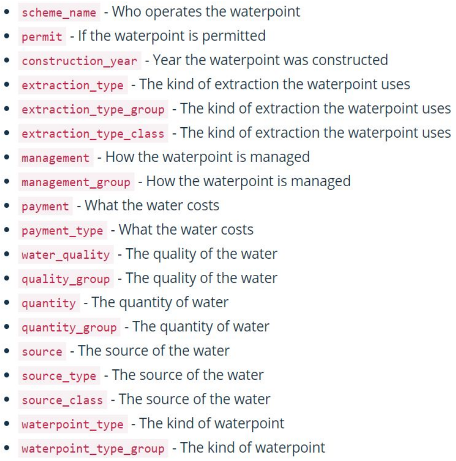
Initial Strategy

1. Get base model (lat, long) model accuracy
2. Intuitive strong predictors besides base model:
    a. Amount TSH, Water Quality, Population, Extraction type, Age
3. Lots of nominal data with high cardinality
    a. Bin these features into something useable
4. Multiple categories with very similar data
5. Multilabel Classification model needed

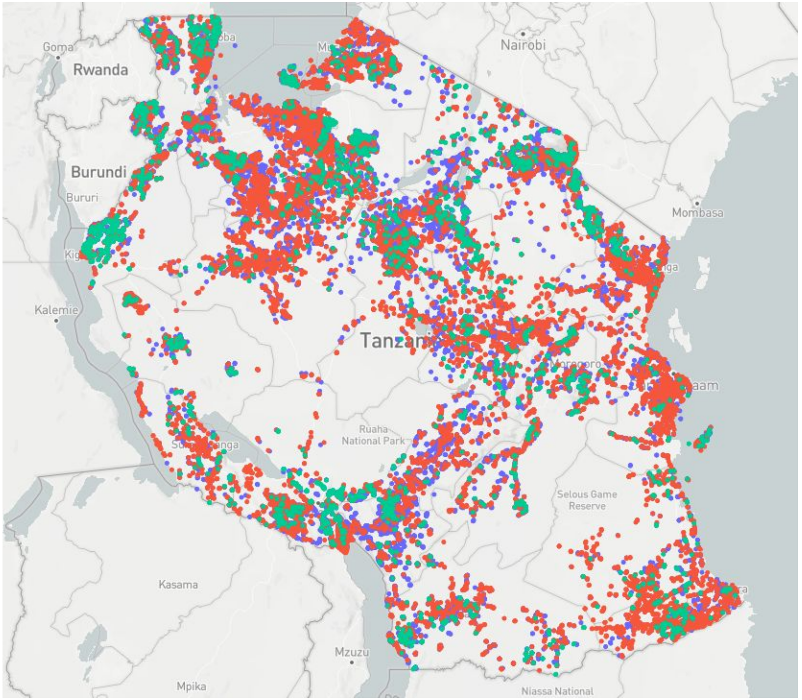


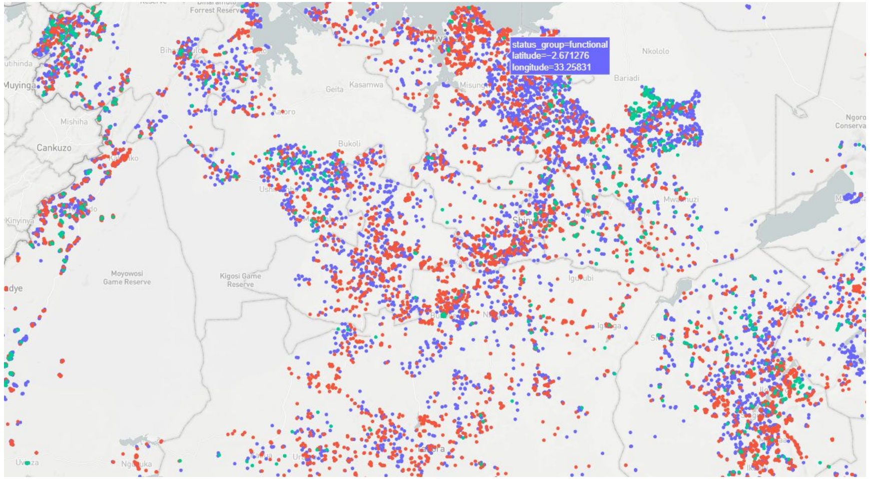
- `amount_tsh` - Total static head (amount water available to waterpoint)
- `date_recorded` - The date the row was entered
- `funder` - Who funded the well
- `gps_height` - Altitude of the well
- `installer` - Organization that installed the well
- `longitude` - GPS coordinate
- `latitude` - GPS coordinate
- `wpt_name` - Name of the waterpoint if there is one
- `num_private` -
- `basin` - Geographic water basin
- `subvillage` - Geographic location
- `region` - Geographic location
- `region_code` - Geographic location (coded)
- `district_code` - Geographic location (coded)
- `lga` - Geographic location
- `ward` - Geographic location
- `population` - Population around the well
- `public_meeting` - True/False
- `recorded_by` - Group entering this row of data
- `scheme_management` - Who operates the waterpoint

59400 observations, 40 features - 28 categorical data - 2 boolean - 10 numerical



- `scheme_name` - Who operates the waterpoint
- `permit` - If the waterpoint is permitted
- `construction_year` - Year the waterpoint was constructed
- `extraction_type` - The kind of extraction the waterpoint uses
- `extraction_type_group` - The kind of extraction the waterpoint uses
- `extraction_type_class` - The kind of extraction the waterpoint uses
- `management` - How the waterpoint is managed
- `management_group` - How the waterpoint is managed
- `payment` - What the water costs
- `payment_type` - What the water costs
- `water_quality` - The quality of the water
- `quality_group` - The quality of the water
- `quantity` - The quantity of water
- `quantity_group` - The quantity of water
- `source` - The source of the water
- `source_type` - The source of the water
- `source_class` - The source of the water
- `waterpoint_type` - The kind of waterpoint
- `waterpoint_type_group` - The kind of waterpoint

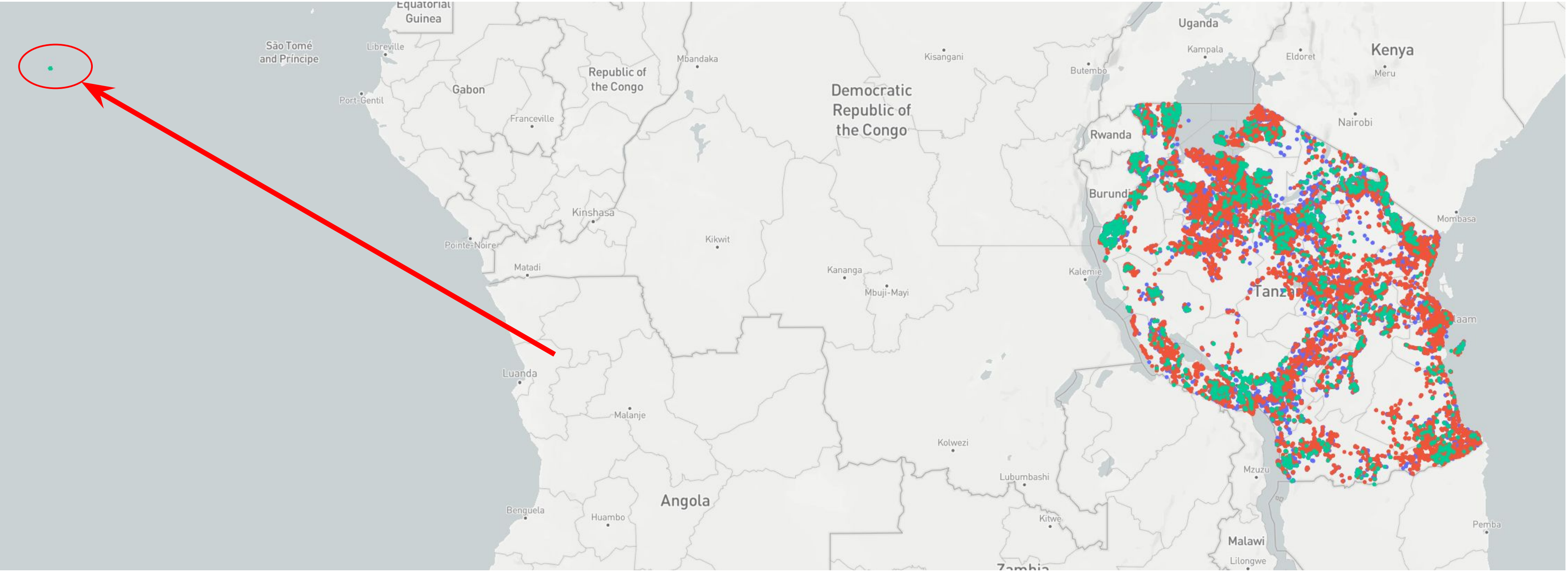




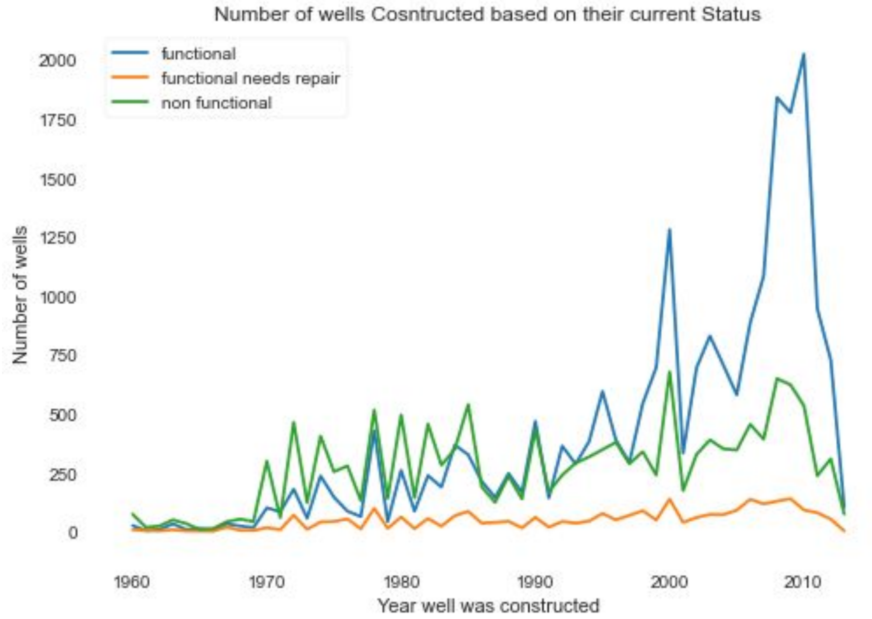
status_group

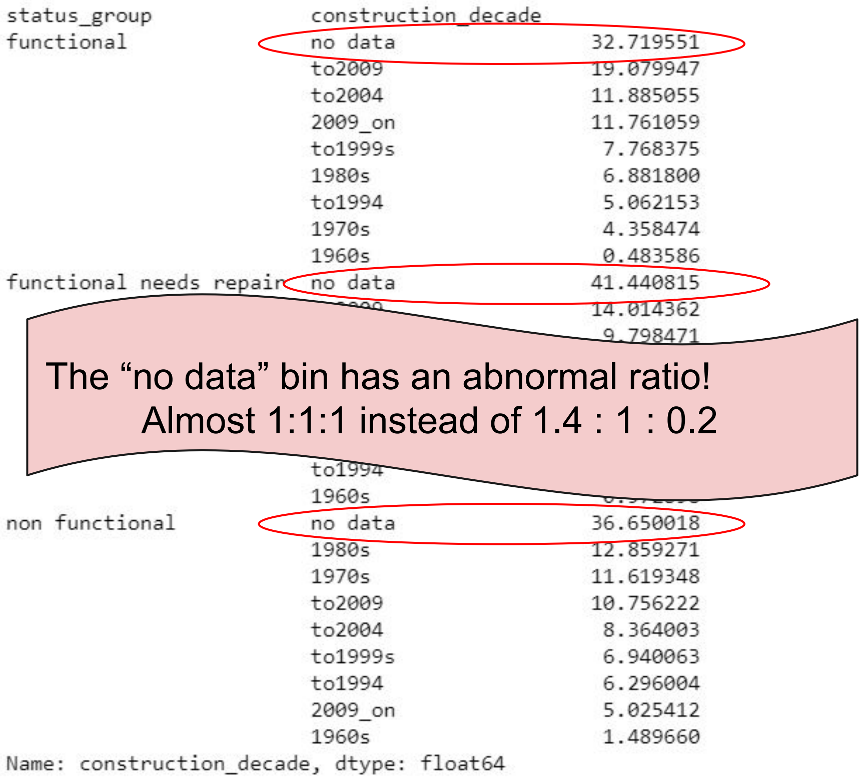
- 🔵 functional
- 🔴 non functional
- 🟢 functional needs repair

Visualization of the data set we received



Some outliers...

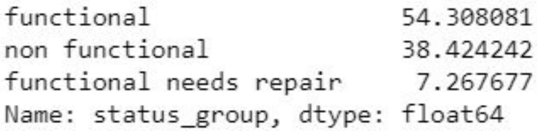




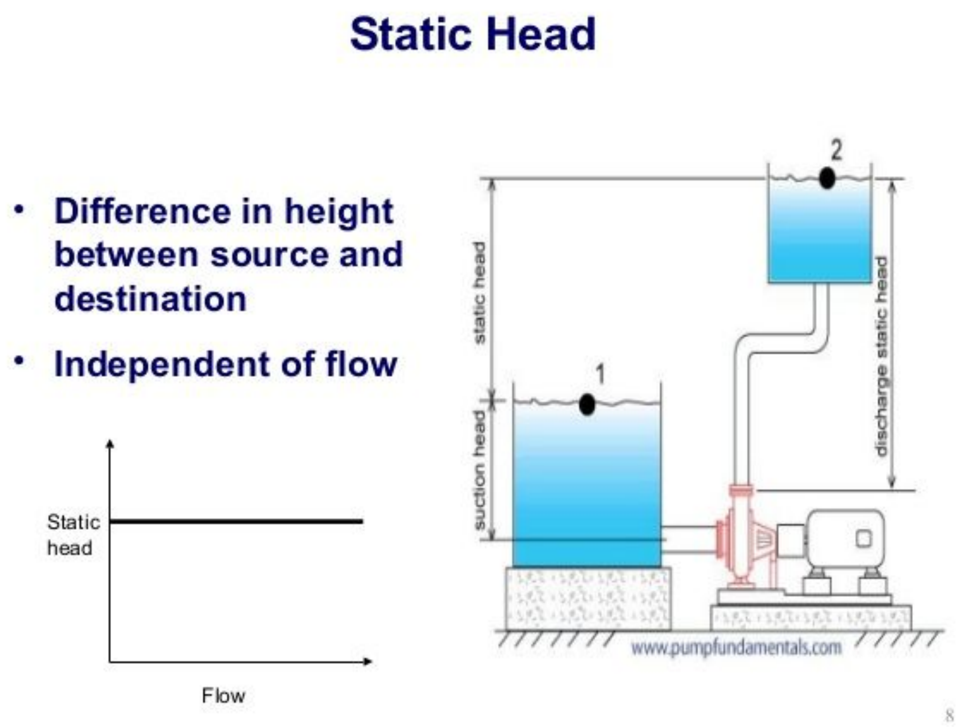
```
status_group           construction_decade
functional             no data                32.719551
                       to2009                 19.079947
                       to2004                 11.885055
                       2009_on                11.761059
                       to1999s                 7.768375
                       1980s                   6.881800
                       to1994                  5.062153
                       1970s                   4.358474
                       1960s                   0.483586
functional needs repair no data               41.440815
                                              14.014362
                                               9.798471
                       to1994
                       1960s
non functional         no data                36.650018
                       1980s                  12.859271
                       1970s                  11.619348
                       to2009                 10.756222
                       to2004                  8.364003
                       to1999s                 6.940063
                       to1994                  6.296004
                       2009_on                 5.025412
                       1960s                   1.489660
Name: construction_decade, dtype: float64
```

The “no data” bin has an abnormal ratio!
Almost 1:1:1 instead of 1.4 : 1 : 0.2

Examples of how I binned features:
Frequency Ratios: 54 : 38 : 7



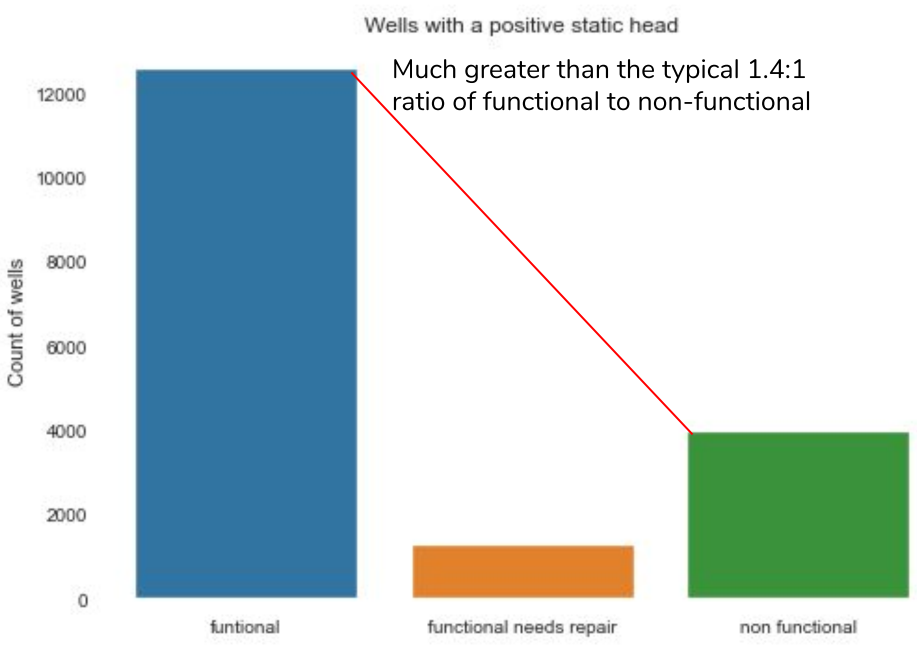
```
functional                 54.308081
non functional             38.424242
functional needs repair     7.267677
Name: status_group, dtype: float64
```



Static Head

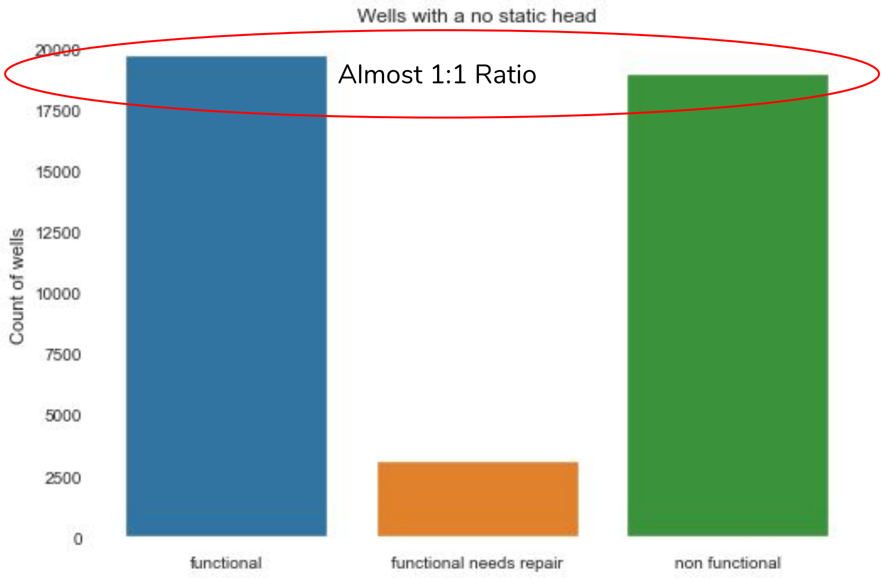
- Difference in height between source and destination
- Independent of flow

What is static head?



Wells with a positive static head

Much greater than the typical 1.4:1 ratio of functional to non-functional

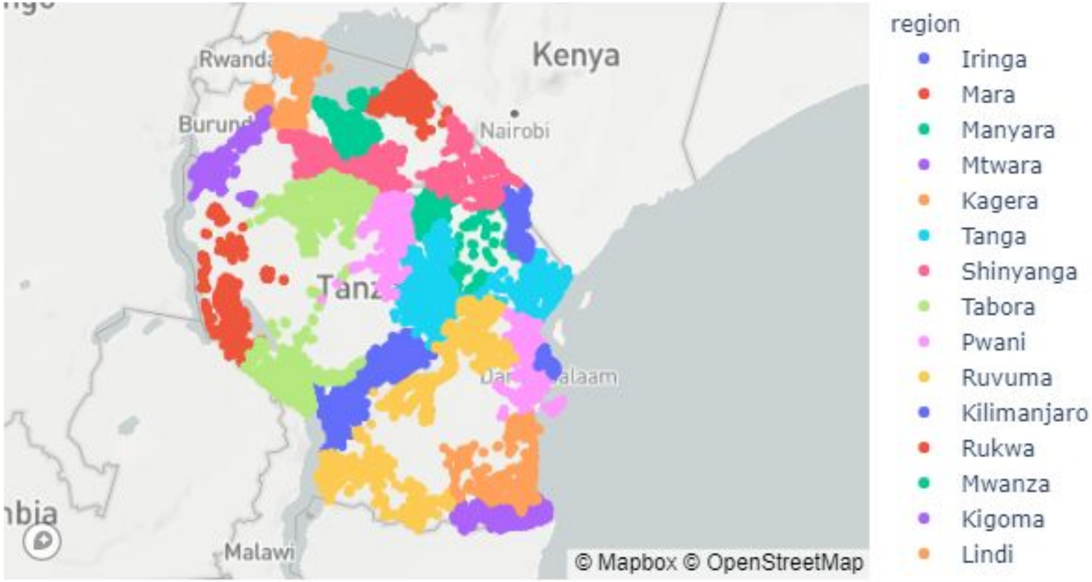


Wells with a no static head

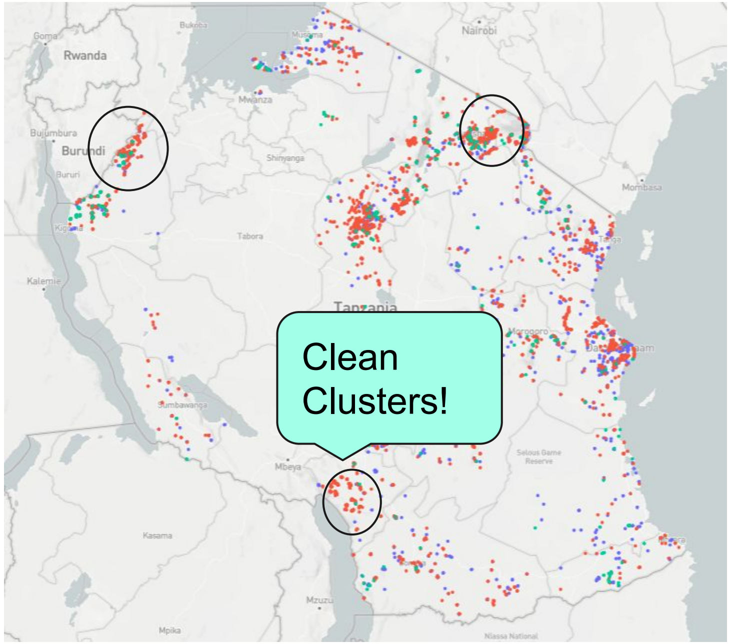
Almost 1:1 Ratio

Total Static Head: Examples of how I binned some features

More Examples of how I used visualizations



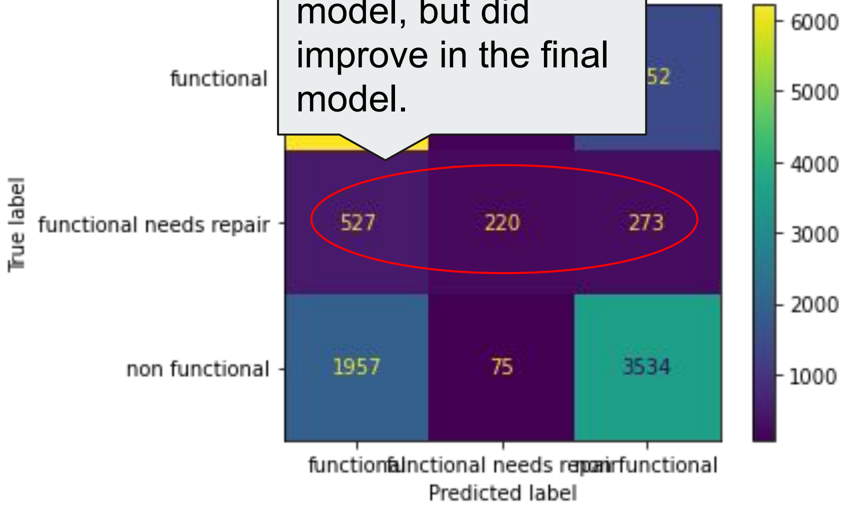
Map of Wells Designated by Region



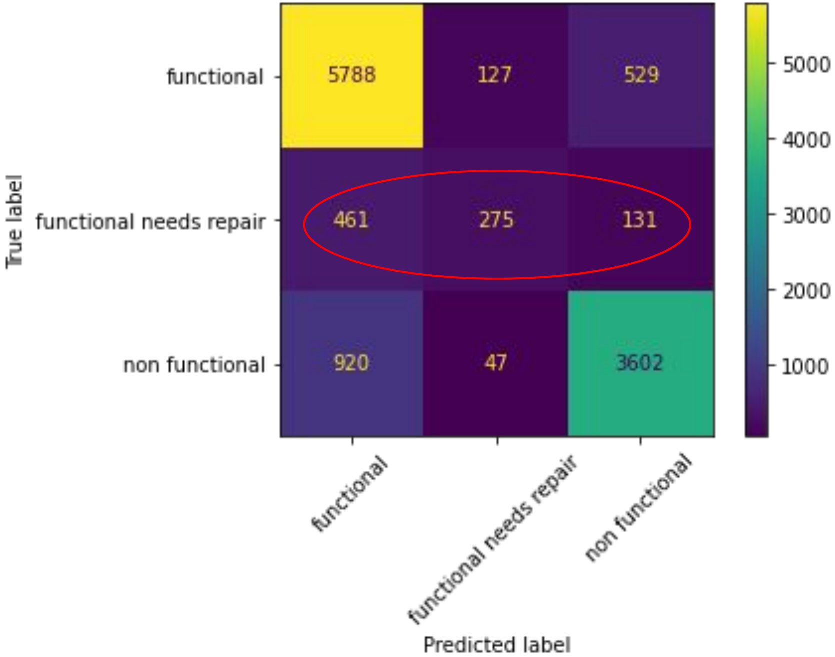
Clean Clusters!

All Wells built 2000-2004

Not great in either model, but did improve in the final model.

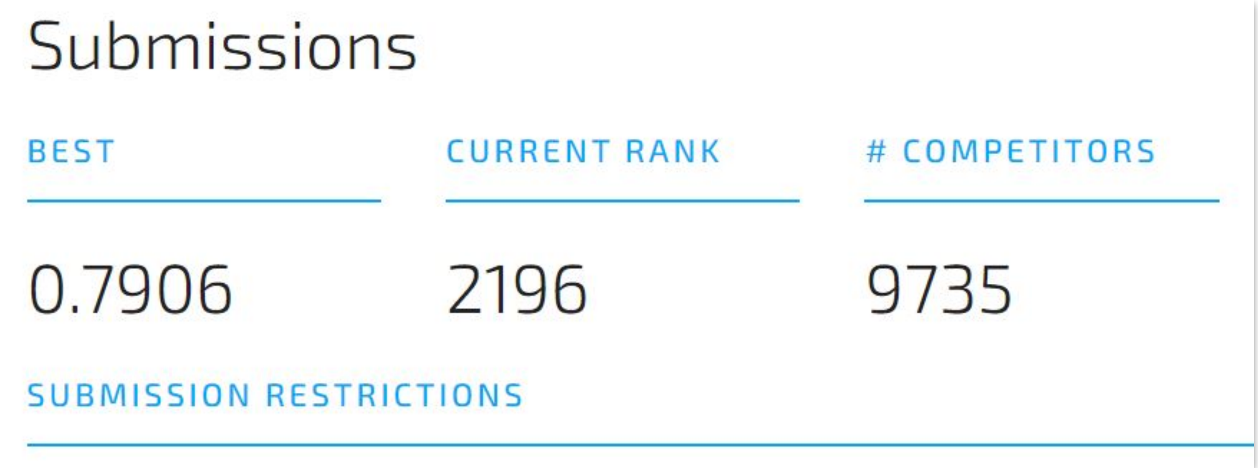


67% Accuracy for the Base Model of Lat/Long



80% Accuracy for Improved Model

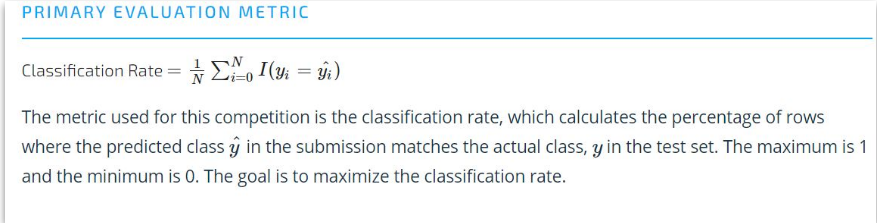
Visualizations of Model Output



Submissions

BEST	CURRENT RANK	# COMPETITORS
0.7906	2196	9735

SUBMISSION RESTRICTIONS



PRIMARY EVALUATION METRIC

Classification Rate $= \frac{1}{N}\sum_{i=0}^{N} I(y_i = \hat{y_i})$

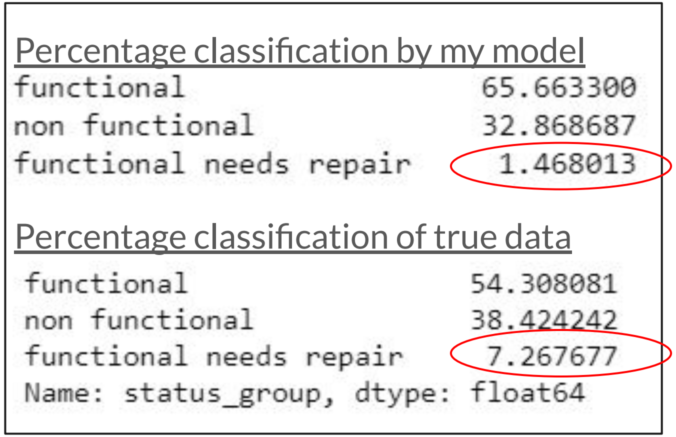
The metric used for this competition is the classification rate, which calculates the percentage of rows where the predicted class $\hat{y}$ in the submission matches the actual class, y in the test set. The maximum is 1 and the minimum is 0. The goal is to maximize the classification rate.

How did my model perform in the contest? *Meh.*

Issues during the project

1. Notebook formatting
    a. Model Train/Test Data
    b. Competition Test Data
2. OneHotEncoding
    a. Sparse Data
    b. Small Bin Numbers
3. Feature Selection
4. Logging Model Information
5. Run time of models

Goal:



Percentage classification by my model

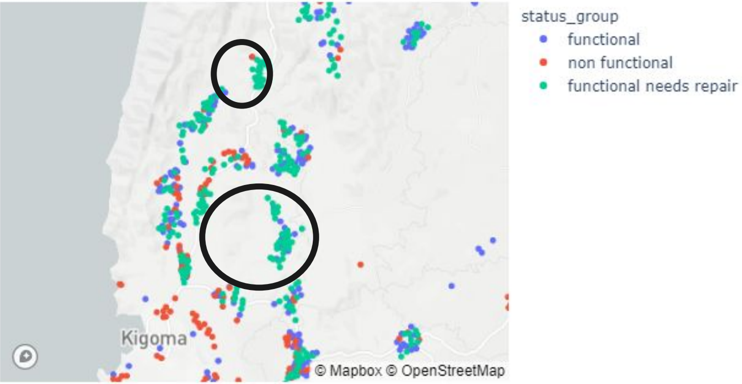
```
functional                   65.663300
non functional               32.868687
functional needs repair       1.468013
```

Percentage classification of true data

```
functional                  54.308081
non functional              38.424242
functional needs repair      7.267677
Name: status_group, dtype: float64
```

Future Work!

Idea 1 - Perform a KNN label prediction feature column and then run an XGBClassifier



Idea 2 - Run this as a two step binary classifier. The first step is to determine functional or not, and the second step is to determine if it needs repairs.

Thanks!

Please use the GitHub link for further data:
https://github.com/bsamaha/Competition---DrivenData---Pump-It-Up