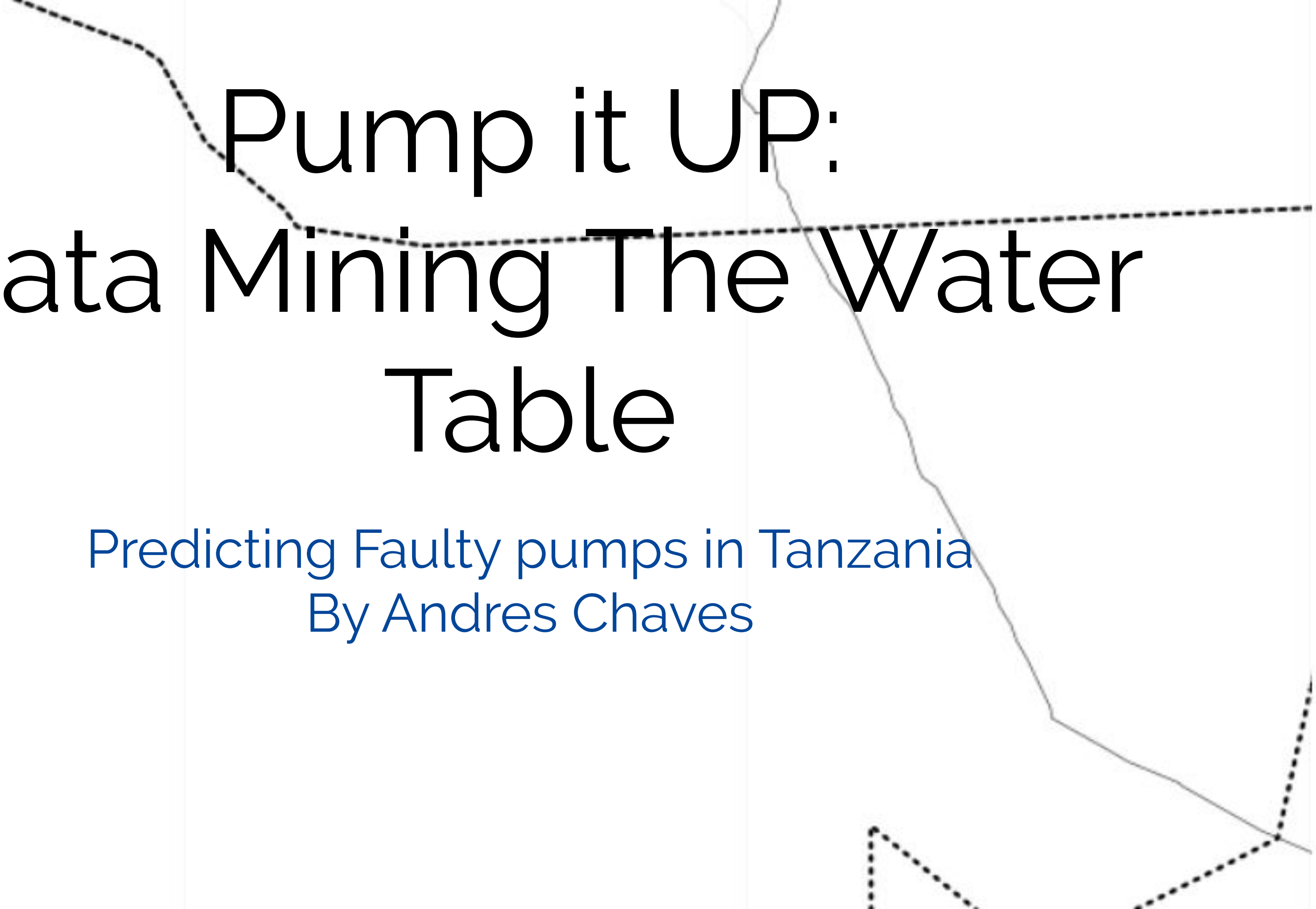


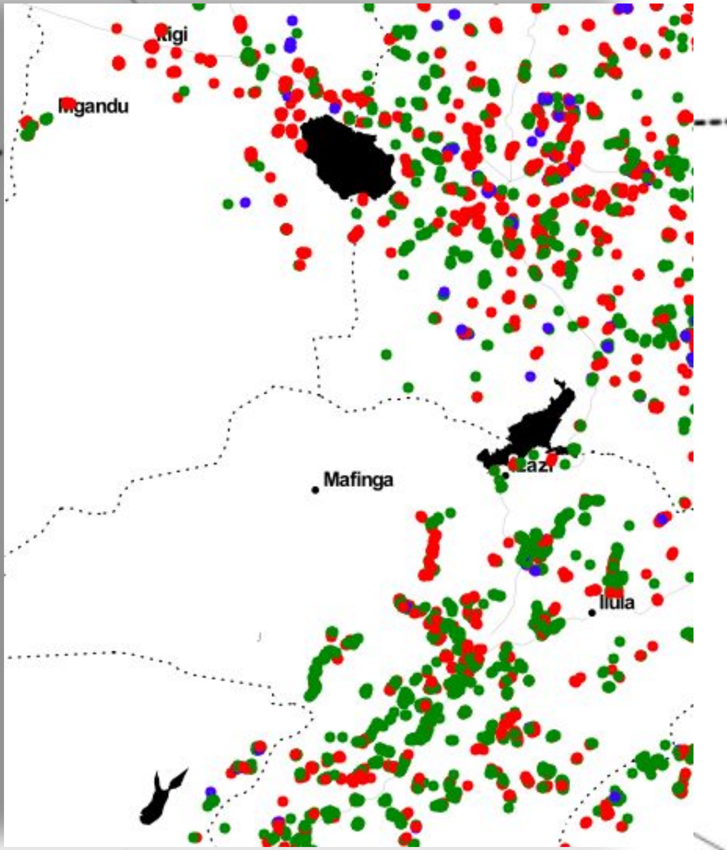
Pump it UP:
Data Mining The Water Table

Predicting Faulty pumps in Tanzania
By Andres Chaves

Motivation:

Review of the model created using sklearn and python for the “Pump It Up” competition sponsored by Driven Data.

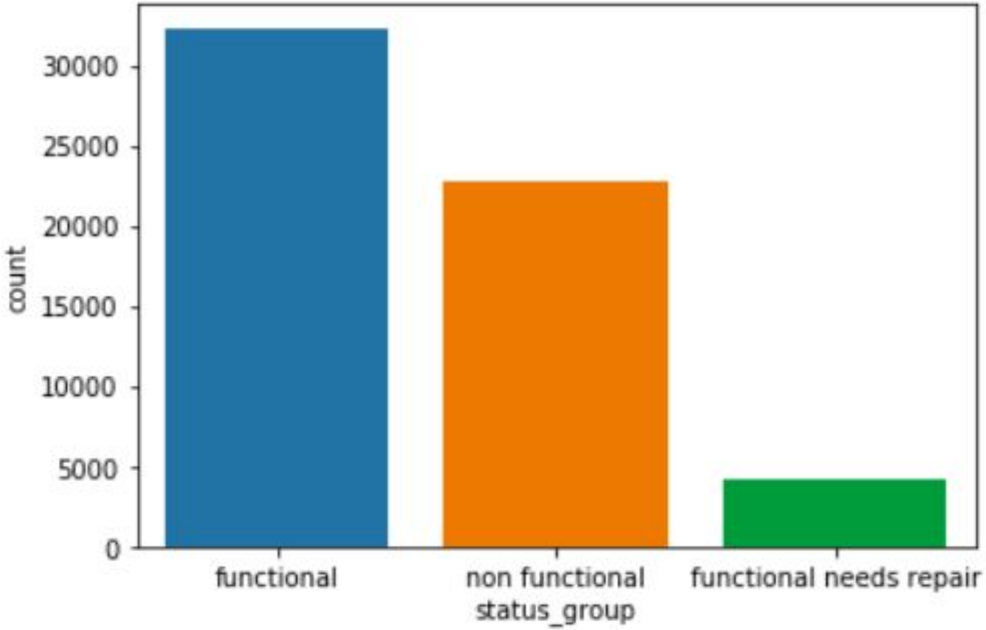
- A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.



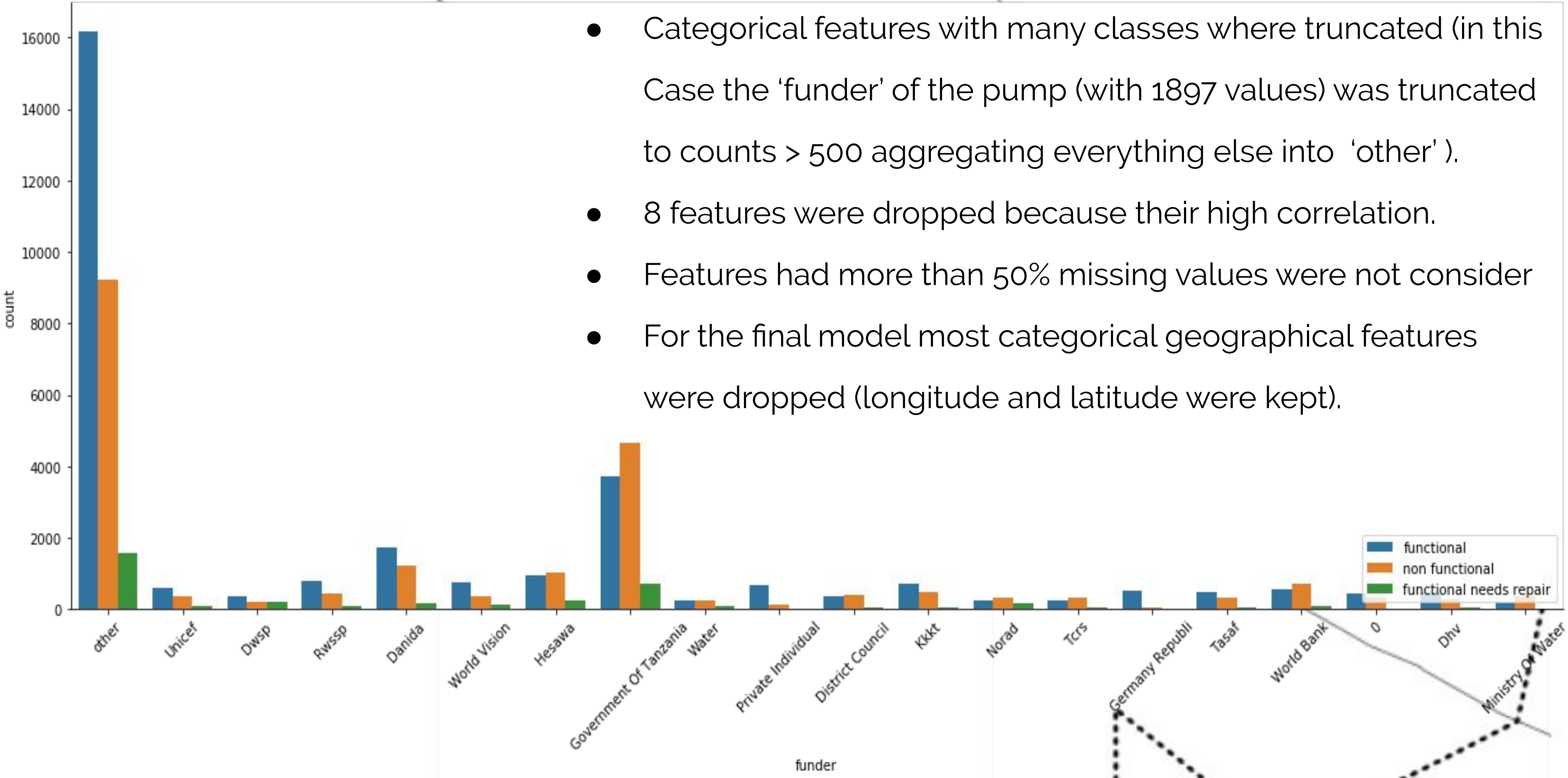
The Goal:

To predict the working status of water pump in Tanzania. This is a multi classification problem with three classes.

1. Functional is the most frequent classwith ~ 54%
2. There is a big class imbalance with the third class
3. Dataset contain 40 posible features (most of the are categorical) with around 59400 observations.
4.



Data Exploration:



- Categorical features with many classes where truncated (in this Case the 'funder' of the pump (with 1897 values) was truncated to counts > 500 aggregating everything else into 'other').
- 8 features were dropped because their high correlation.
- Features had more than 50% missing values were not consider
- For the final model most categorical geographical features were dropped (longitude and latitude were kept).

Base line

In order to set a baseline two models were considered:

- One that returned the most frequent class (54%)
- And a Knn model with only latitude and longitude as features

```
Testing F1 Score: 0.7279239849945974
Testing Accuracy Score: 0.7390572390572391
```

	Functional	Need repair	Non functional
Functional	6629	112	1267
Needs repair	2340	1147	243
Non Functional	2452	92	3236

The Model: Random Forest

The final model was found with grid searchCV using a random forest classifier

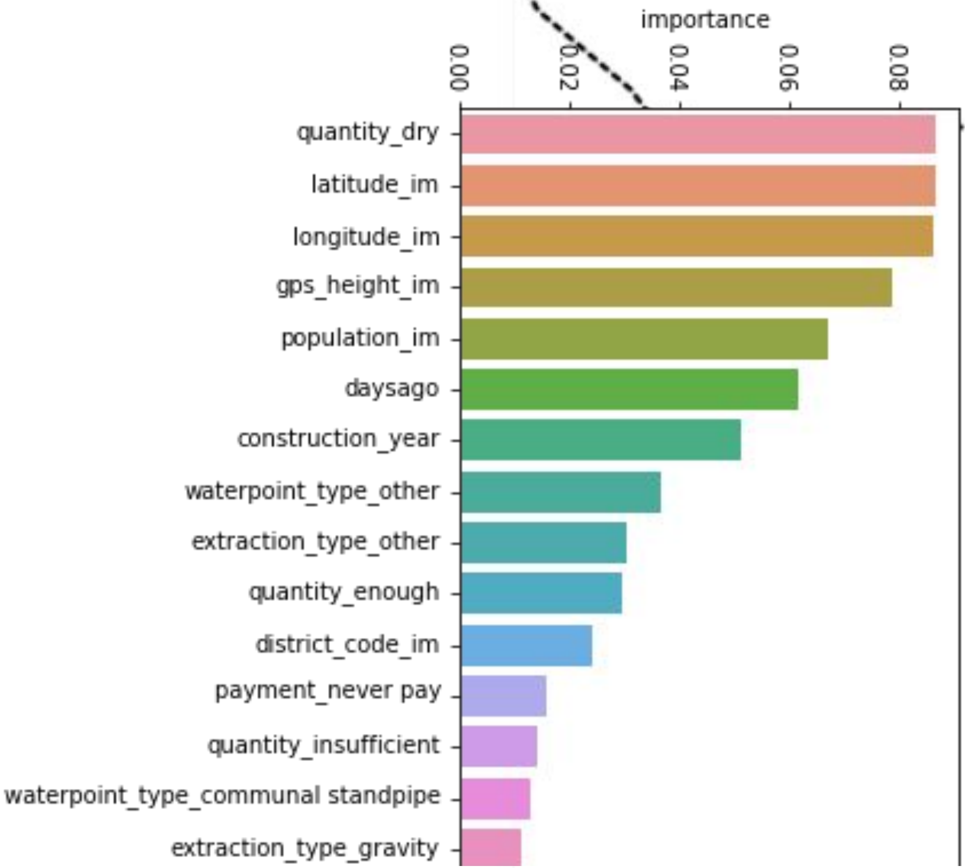
- Generated a 10% improvement of KNN
- Best Model scores:

```
Training F1 Score: 0.951861830971213
Training Accuracy Score:
0.9529741863075196 Testing F1 Score:
0.8124166310447034        Testing Accuracy
Score: 0.8189225589225589
```

- Trying to address class imbalance I did on run with `class_weight='balanced'`. It gave me a worst model.

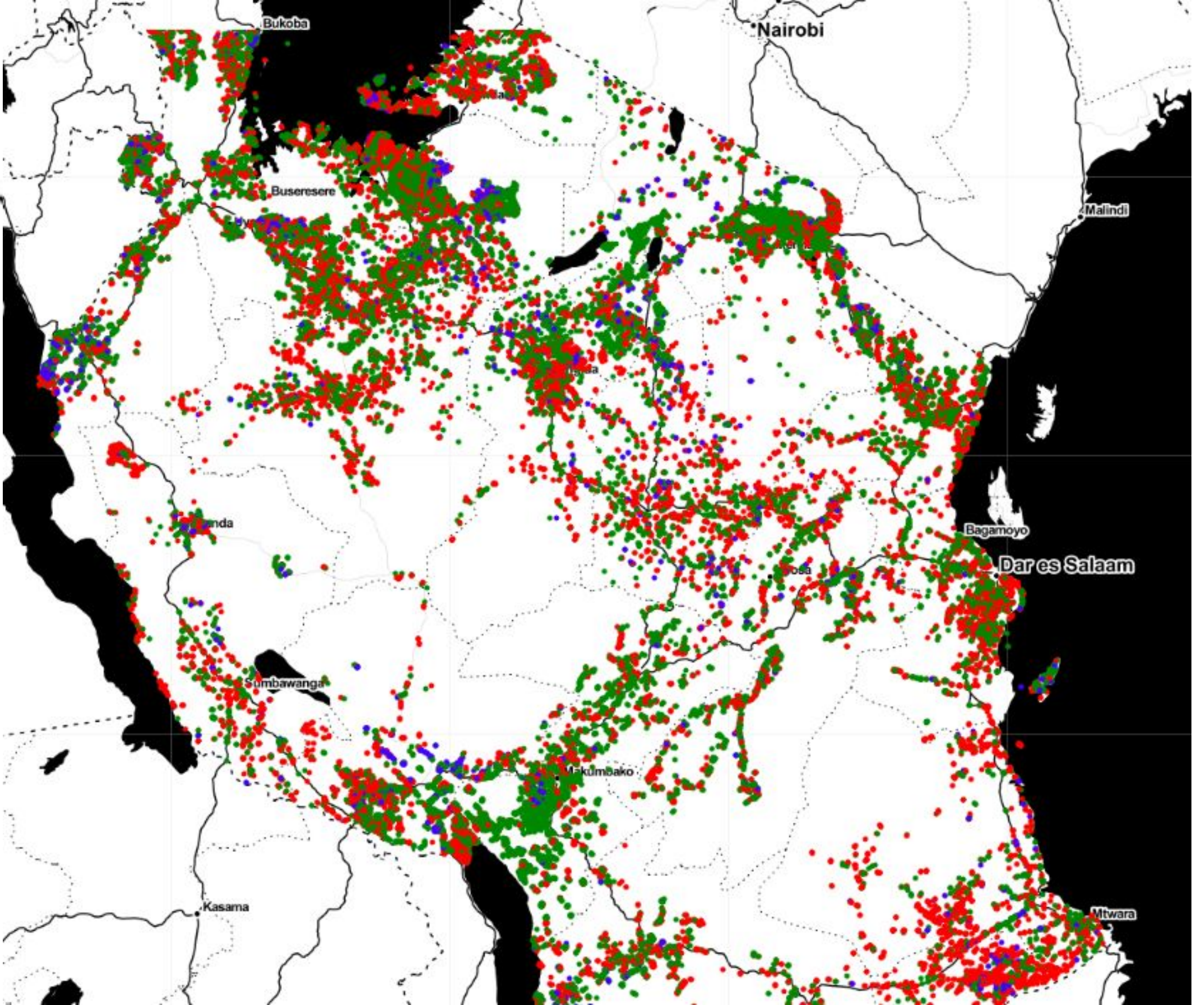
	Functional	Need repair	Non functional
Functional	7185	179	644
Needs repair	534	368	160
Non Functional	1091	370	4608

Feature importance:



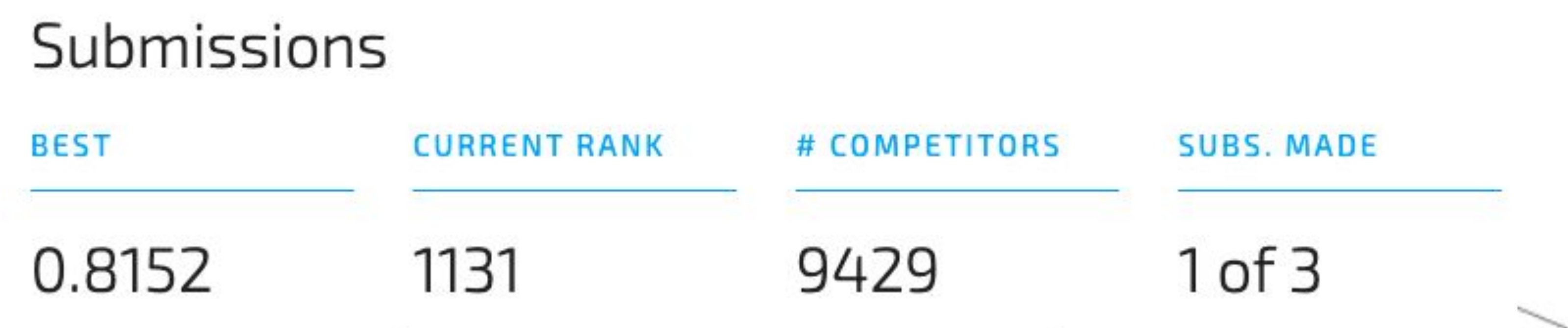
quantity_dry	0.08657181892
latitude_im	0.08647516456
longitude_im	0.08584935745
gps_height_im	0.0786328644
population_im	0.06683669854
daysago	0.06119725154
construction_year	0.05078358004
waterpoint_type_other	0.03631512461
extraction_type_other	0.02989290728
quantity_enough	0.0291166562
district_code_im	0.02372770646
payment_never pay	0.01546280687
quantity_insufficient	0.0137755552

- Quantity of water is the top feature. suggesting that many pumps could be just dry
- Longitude and Latitude are at the top of my Feature importance
- Tanzania is the country in Africa with the biggest range of altitude (hight is the 4th important feature)
- How long ago the observation was taken ‘daysago’ is also important



Food for thought:

- More feature engineering is needed. Increase subject knowledge
    - (longitude and latitude)
    - Socio economic data
- Try more things
    - Run XGboost Model
    - Add voting



Submissions

BEST	CURRENT RANK	# COMPETITORS	SUBS. MADE
0.8152	1131	9429	1 of 3