# Pump It Up

## Data Mining the Water Table

by Kourosh Alizadeh

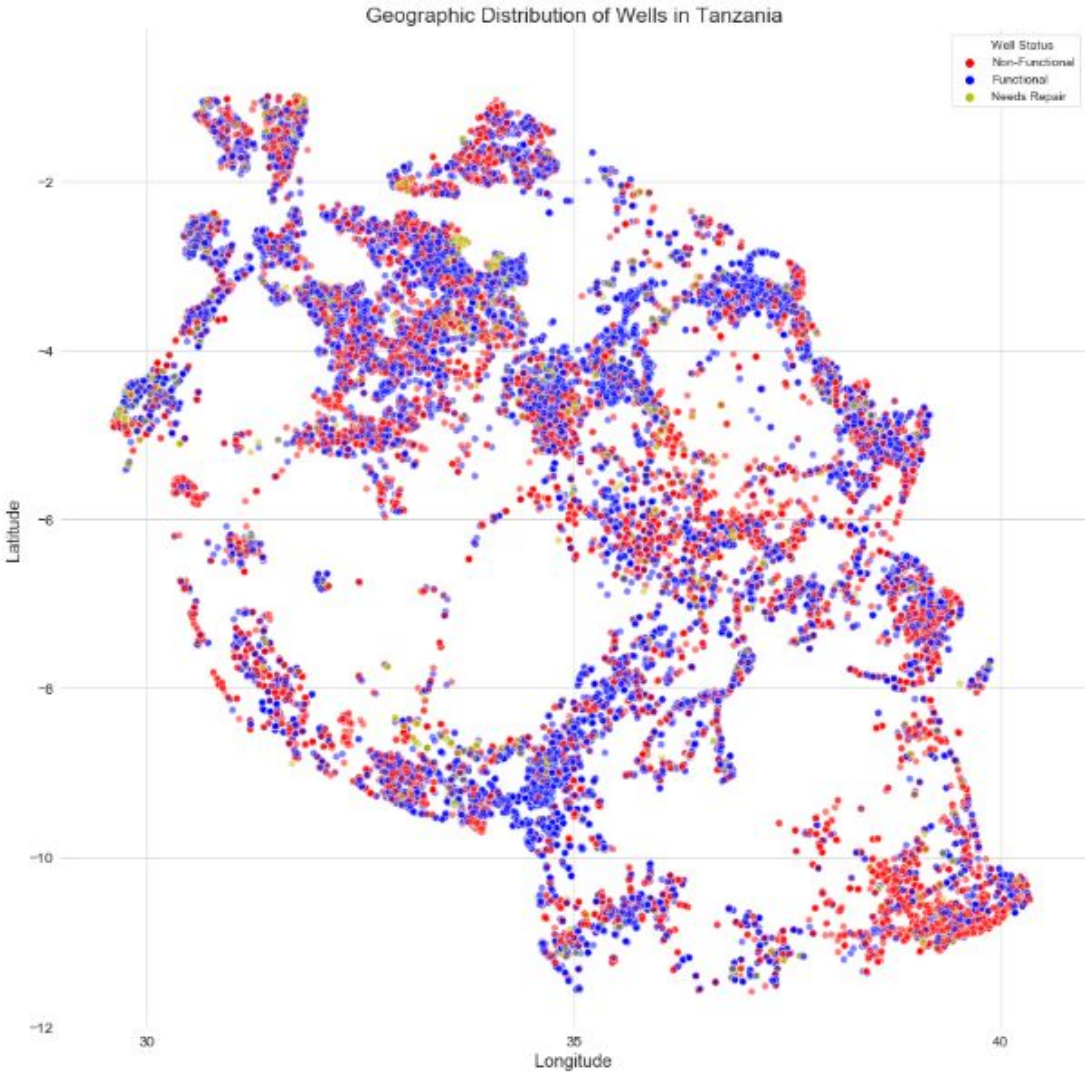
# Project Goals & Methodology

- Our goal is to predict the functionality of wells in Tanzania

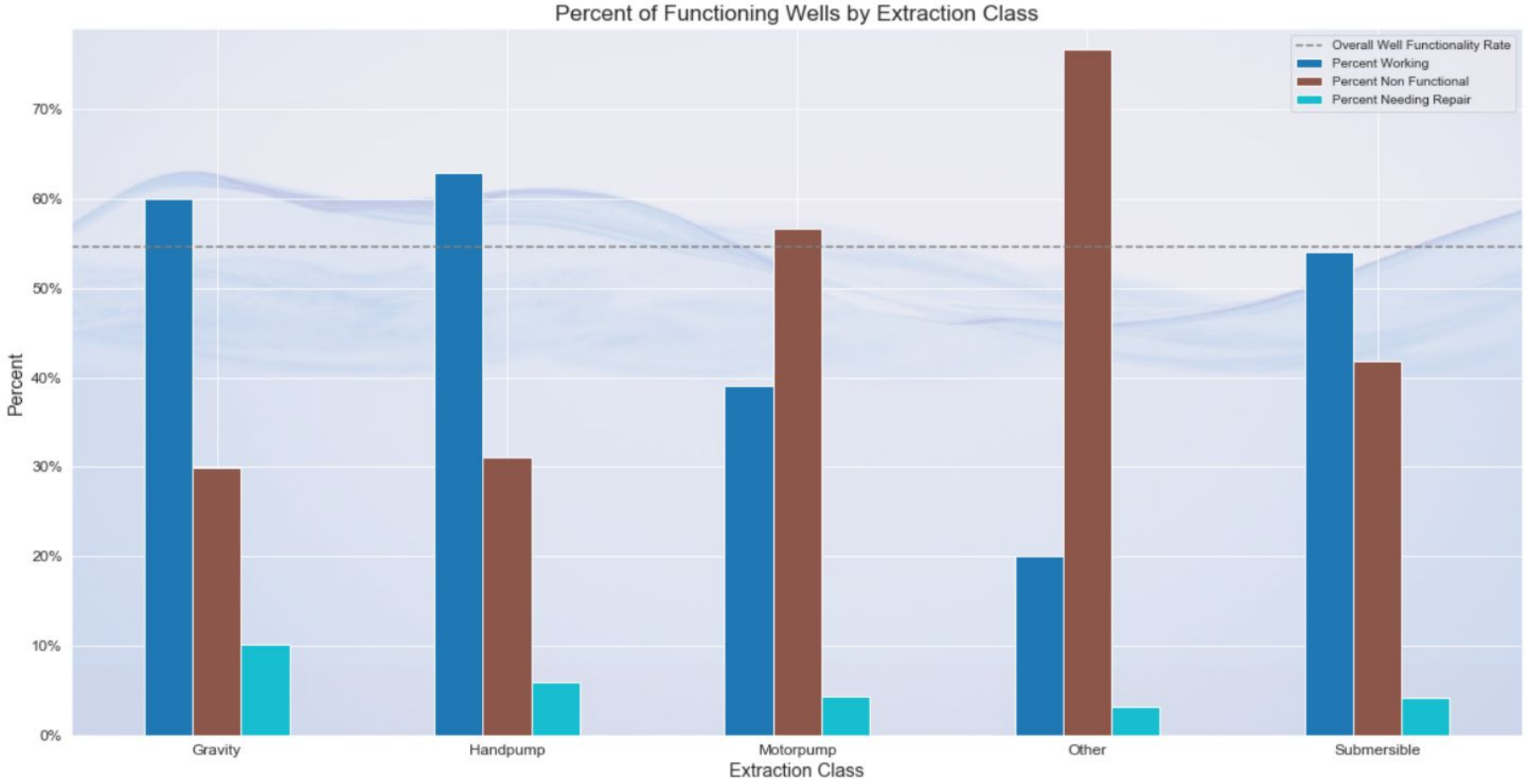
- We used the **OSEMN** methodology.
  - **Obtain** - data provided by DrivenData as part of an online data science competition
  - **Scrub** - removed null values, imputed values to various place-holders
  - **Explore** - built a function to visualize the connection between features and well functionality.
    - 5 EDA questions examined in more detail
  - **Model & Interpret** - built a random forest model and examined feature importance

Q1 - Is there any geographic pattern for functioning wells?

Some pattern. Wells tend to cluster in functionality groups, and more fail in the south-east

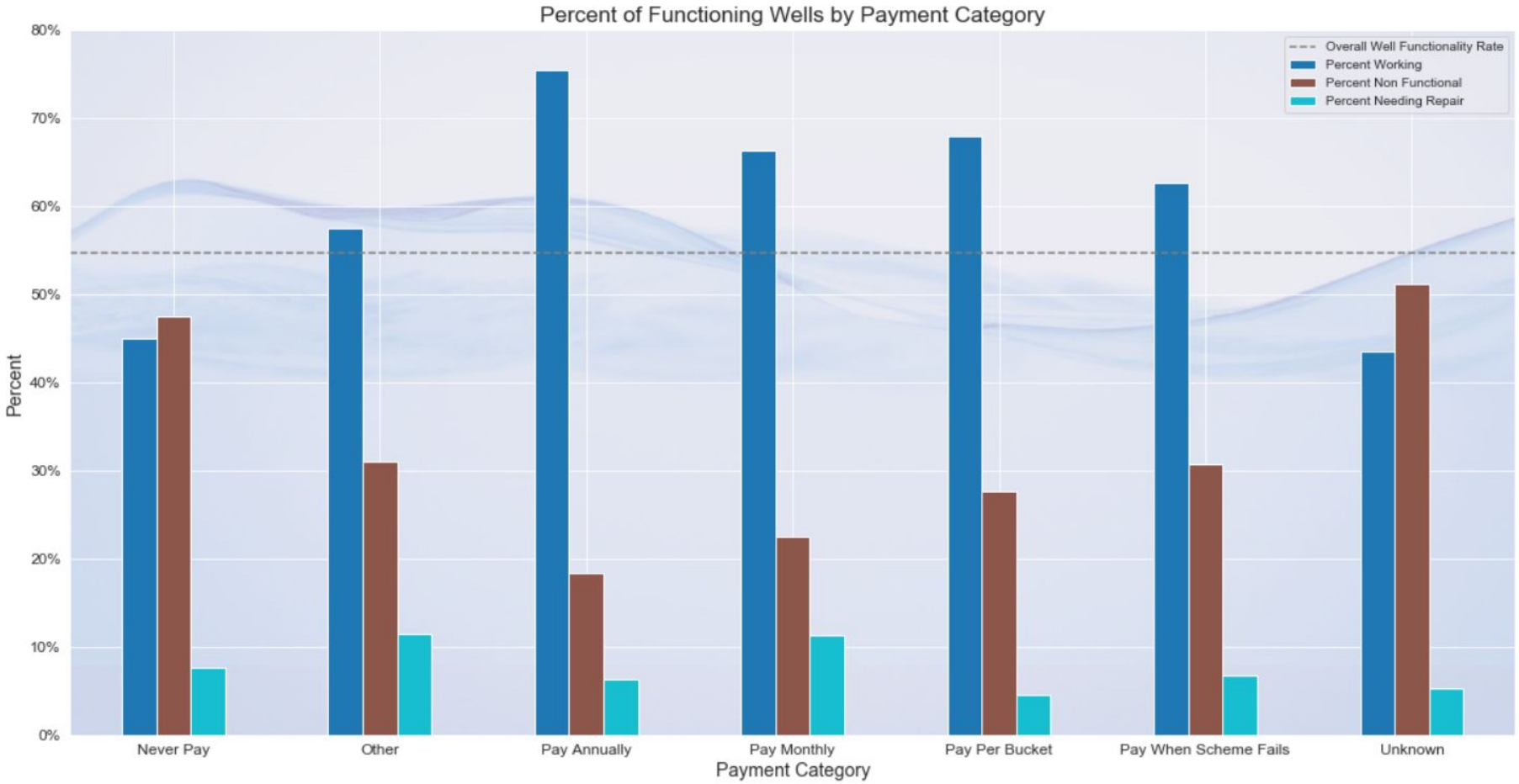


# Q2 - Does Functionality vary w/ Extraction Method?

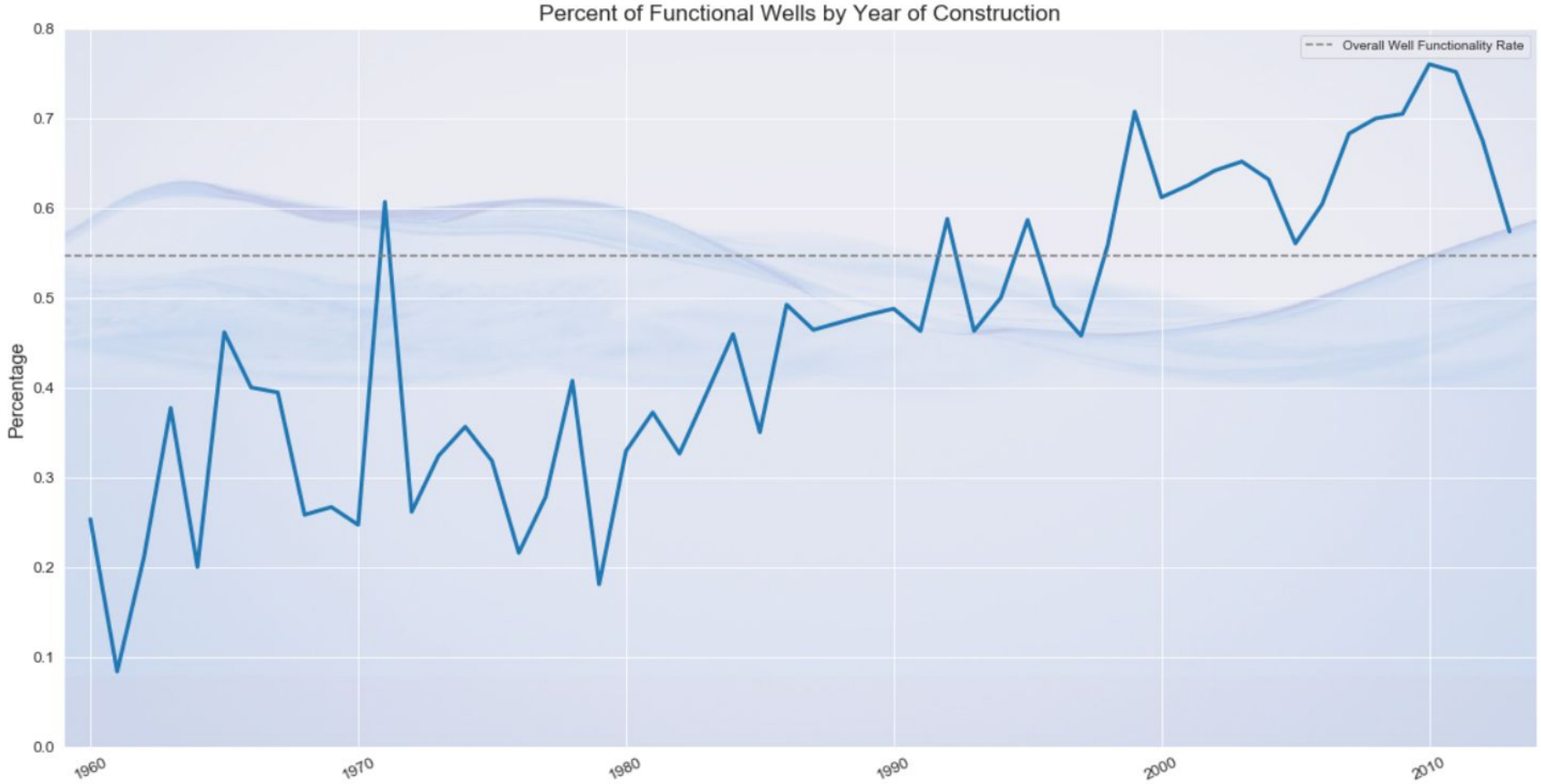


# Q3 - Does Functionality vary w/ Waterpoint Type?

# Q4 - Does Functionality vary w/ Payment?



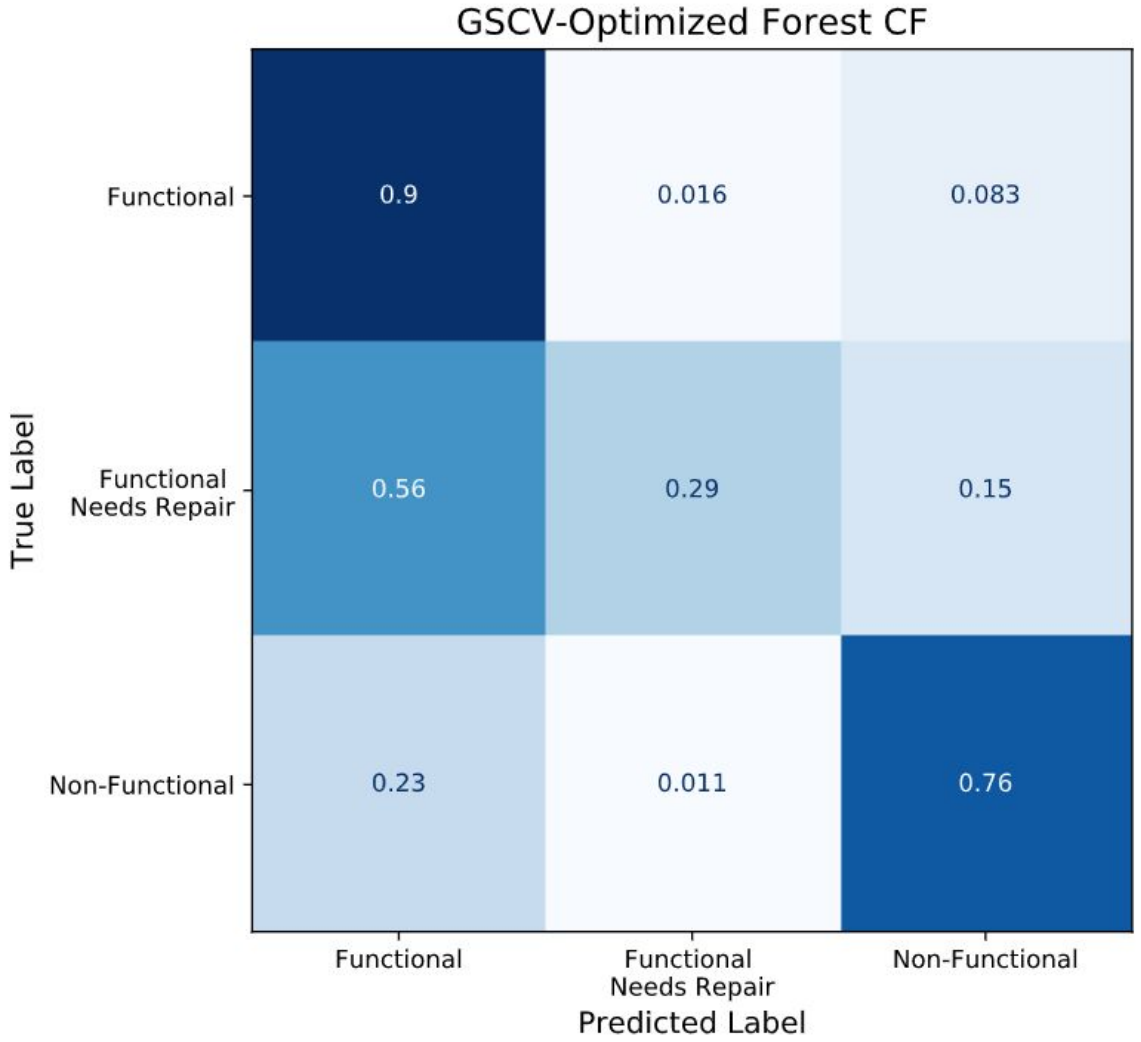
# Q5 - Does Functionality vary w/ Age?



# Modelling the data

We built a Random Forest classifier and optimized it for accuracy.

It does a solid job of classifying functional wells correctly, but makes more mistakes with non-functional wells. Those in need of repair are often misclassified.

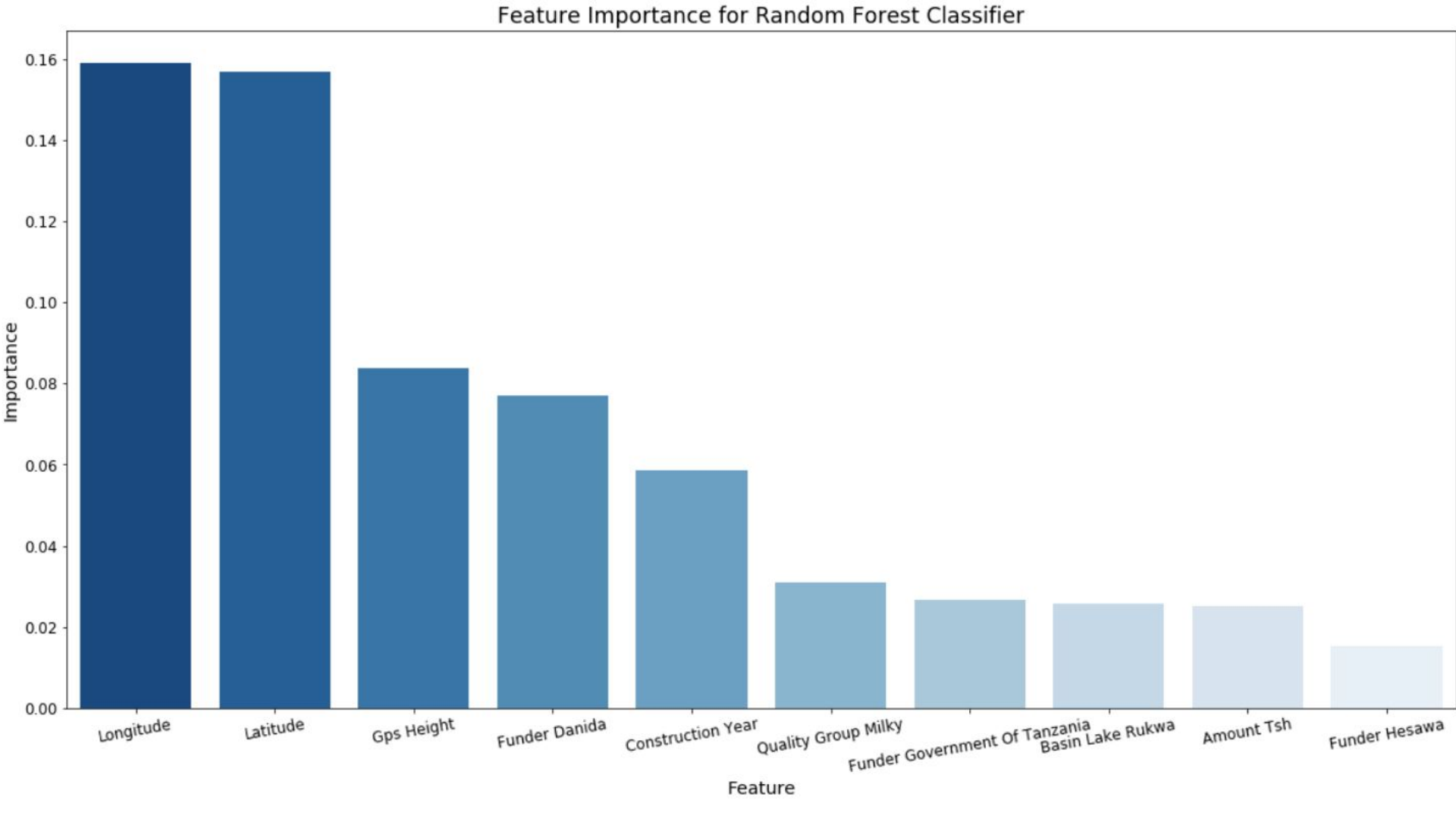


# Overall Submission Score

| BEST | CURRENT RANK | # COMPETITORS |
| --- | --- | --- |
| 0.8071 | 1767 | 10296 |

Our model scored .8071 accuracy on the competition site.

We are in the 83rd percentile of scores at the time of submission.


Feature Importance for Random Forest Classifier
Importance
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Longitude
Latitude
Gps Height
Funder Danida
Construction Year
Quality Group Milky
Funder Government Of Tanzania
Basin Lake Rukwa
Amount Tsh
Funder Hesawa
Feature

# Interpret Data - Recommended Actions

- The primary predictive features were geographic; let's identify problem regions and direct funding to well-building entities in those regions
- Focus on gravity and hand-pump extraction methods
- Charge for well use - annual payment plans are optimal
- Give contracts to successful funding agencies like Danida

# Future Analysis

- Find features that uniquely identify wells in need of repair; these should be a top priority

- Build a model for water scarcity as opposed to well functionality
  - Many broken wells exist near functioning wells; the number of broken wells does not straightforwardly indicate need
  - Instead, a metric like functioning wells within a certain distance might be a better metric to use
    - Especially if we incorporate population data

- Combine models for greater accuracy

# Thank you!