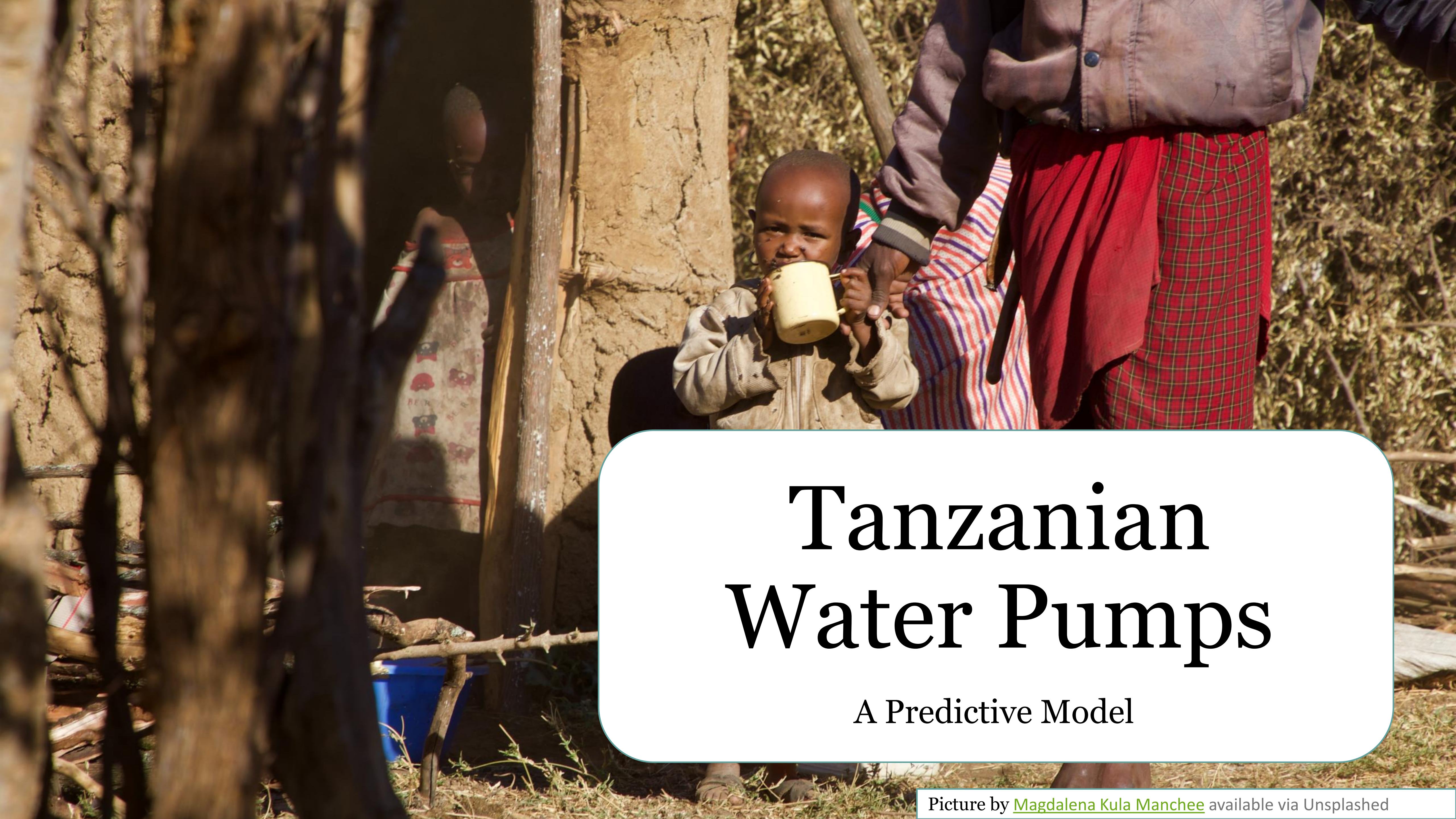


Tanzanian Water Pumps

A Predictive Model

Picture by Magdalena Kula Manchee available via Unsplashed

Initial Problem

There are significant NGO resources seeking to help improve and maintain Tanzania’s water infrastructure.

This analysis will address two questions:

- What pumps are likely to be broken or in need of repair?
- What features of the pumps are most predictive of pump failure?

Method

- Produce a classification models to predict pump functioning

- Using the feature importance of those models to infer what affects pump status

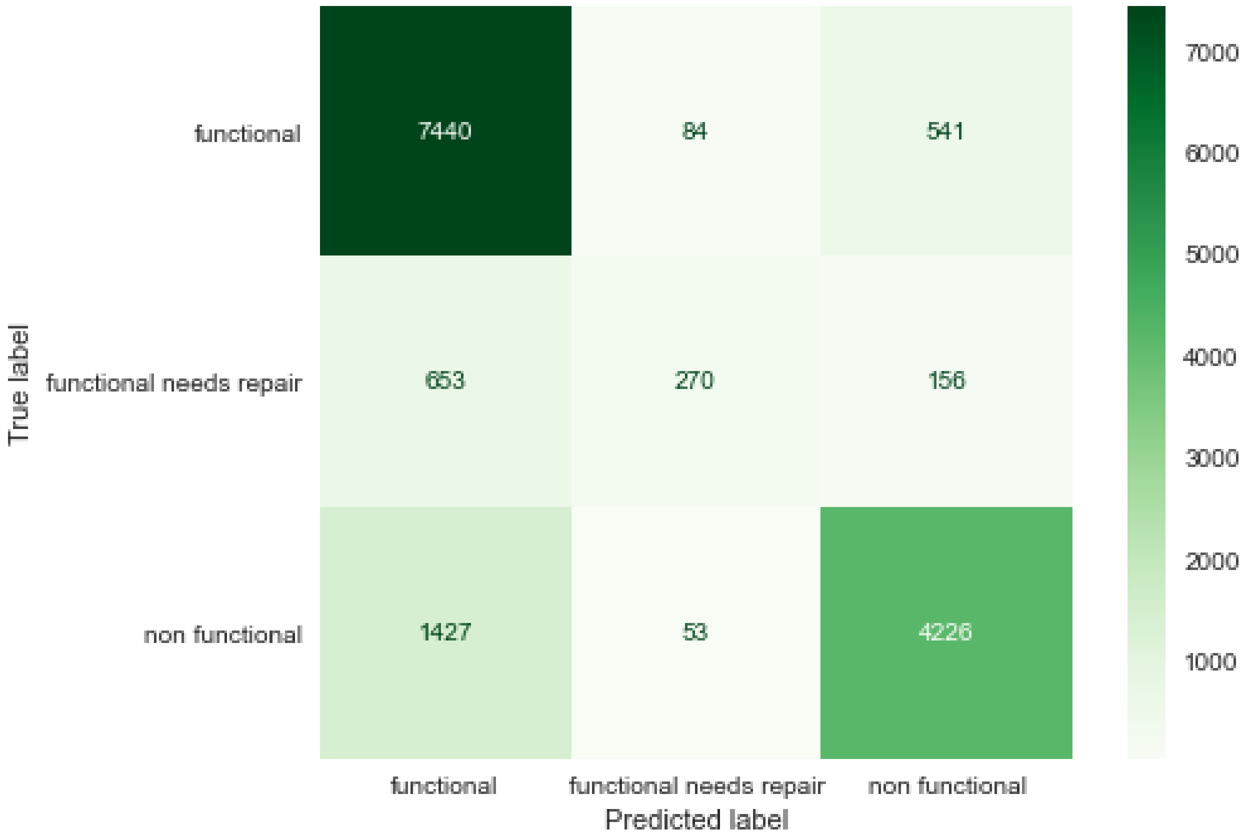
- Multiple modeling methods were used, including, logistic, random forest, XGBoost and CatBoost classifiers

Available Data

- Data was taken from the Pump It Up competition website.
- Data was gathered by the Tanzanian government from 2013-2016.
- The raw training data consisted of 59400 observed pumps with 39 recorded features beyond the id number.

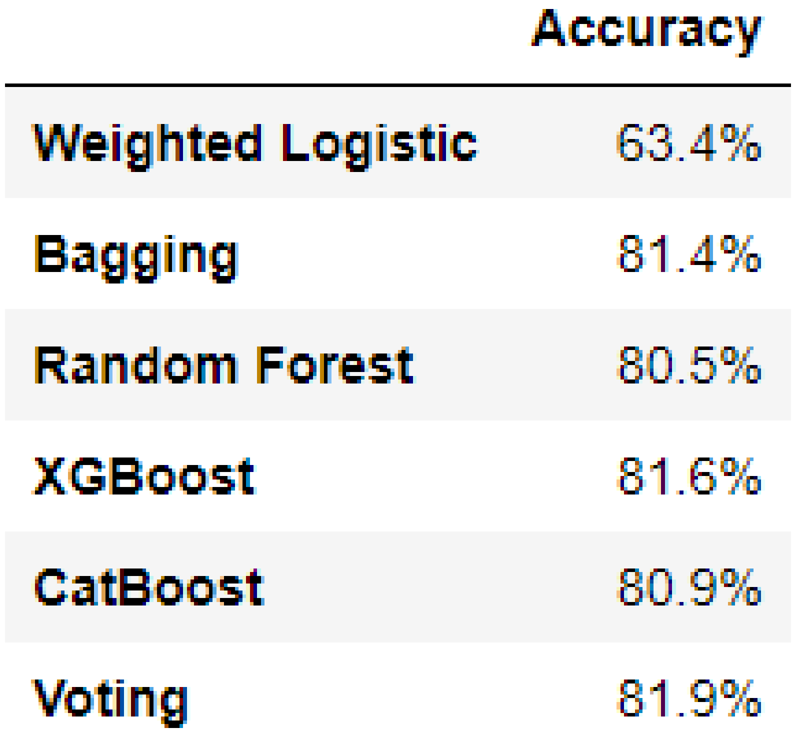
Example Model Results

An untuned XGBoost model had an accuracy of 79.9%



Overall Results

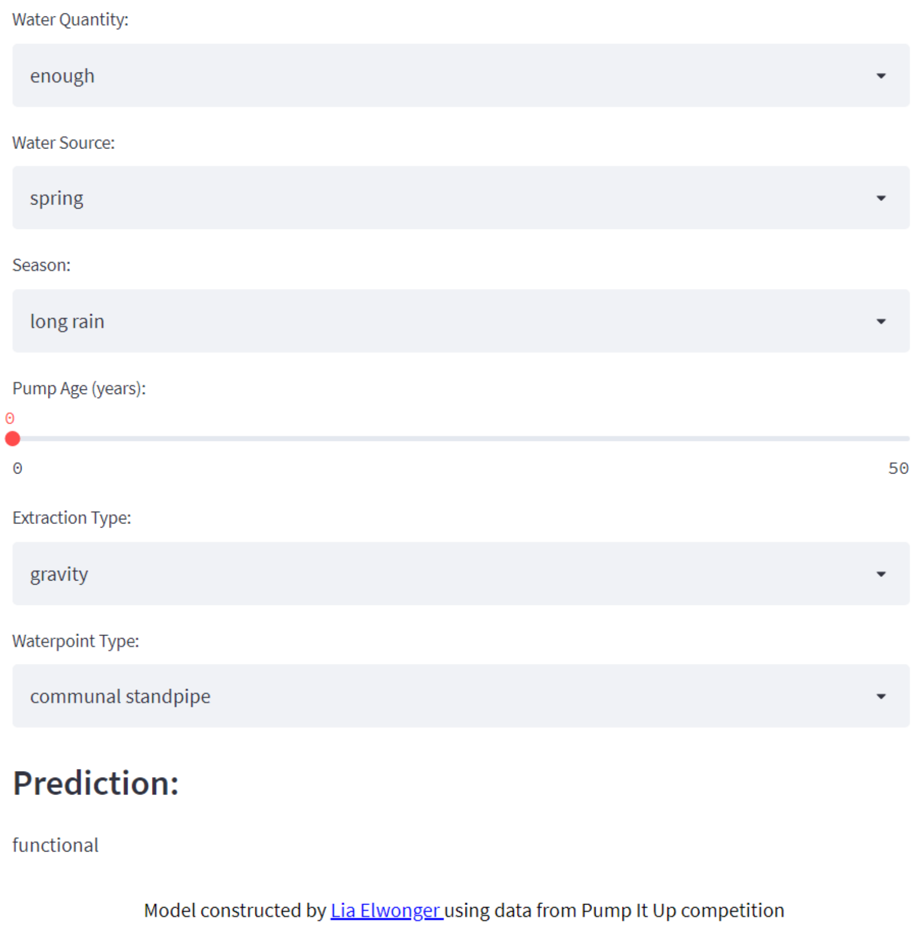
The models were all tuned leading to these final performance results for each type of model:



	Accuracy
Weighted Logistic	63.4%
Bagging	81.4%
Random Forest	80.5%
XGBoost	81.6%
CatBoost	80.9%
Voting	81.9%

Prediction App

The models XGBoost model is small enough to be made into a simple app that you can find on the Streamlit website.



Water Quantity:

enough

Water Source:

spring

Season:

long rain

Pump Age (years):

0

0 50

Extraction Type:

gravity

Waterpoint Type:

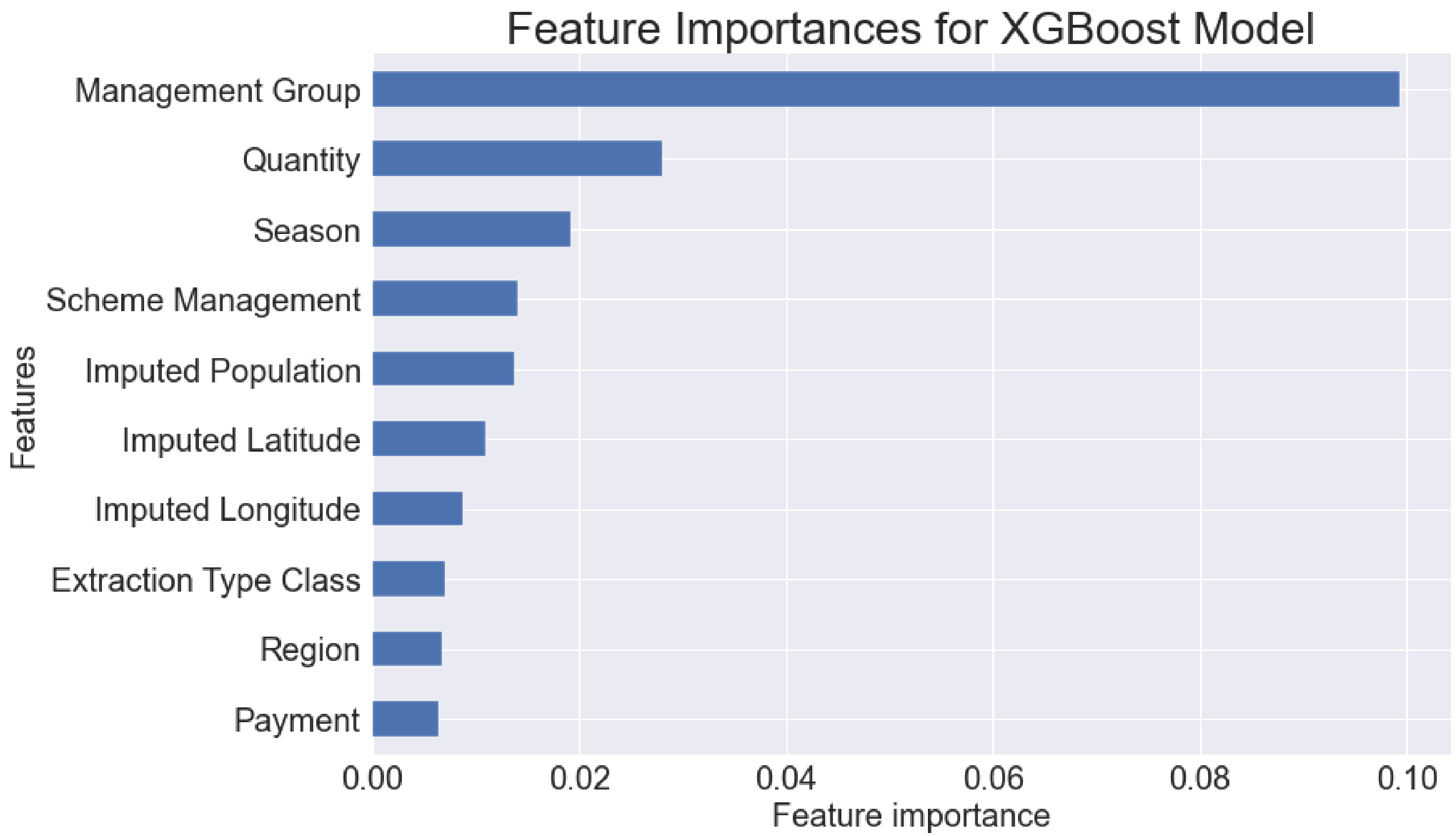
communal standpipe

Prediction:

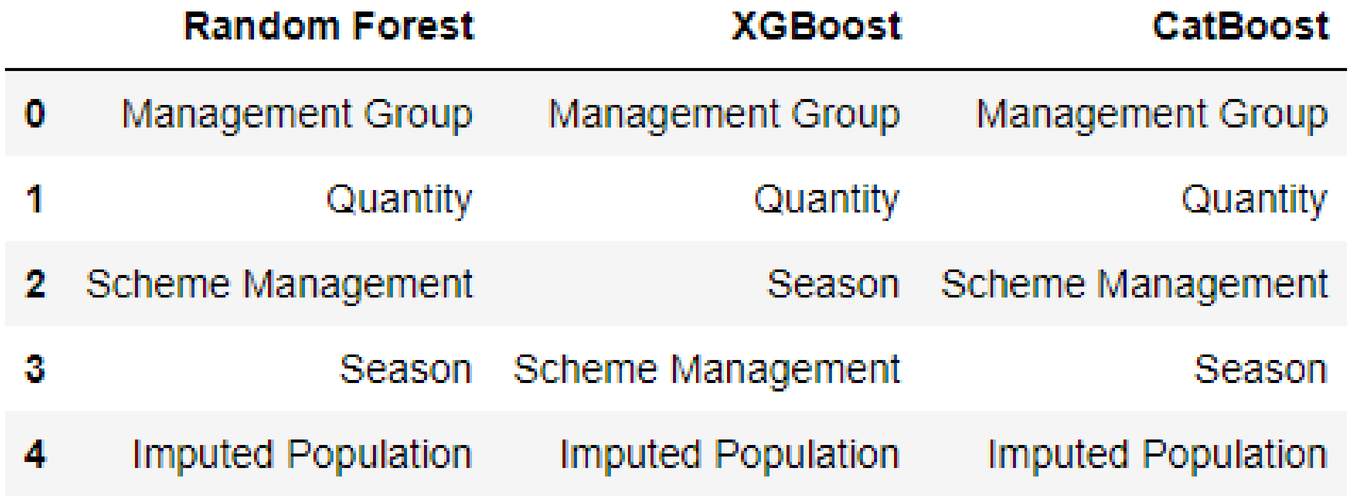
functional

Model constructed by Lia Elwonger using data from Pump It Up competition

Example Feature Importance Results



Top Five Features of Best Performing Models



	Random Forest	XGBoost	CatBoost
0	Management Group	Management Group	Management Group
1	Quantity	Quantity	Quantity
2	Scheme Management	Season	Scheme Management
3	Season	Scheme Management	Season
4	Imputed Population	Imputed Population	Imputed Population

Conclusions

Be very careful in the selection of who will manage your installed pump.

Gather data about the same pumps across seasons, since there is a large seasonal affects in water available.

Use the model to predict what pumps have seasonal variance and provide other sources of water if possible, to these areas.

Limitations

It is important to recognize a categorization model is not a guarantee for causal inference.

For example, it may be that certain managers don’t cause failure, but are given worse pumps.

To get deeper insight a RCT or other form of causal inference would likely be required.