Pump it Up: Data Mining the Water Table

Seth Chart

Data

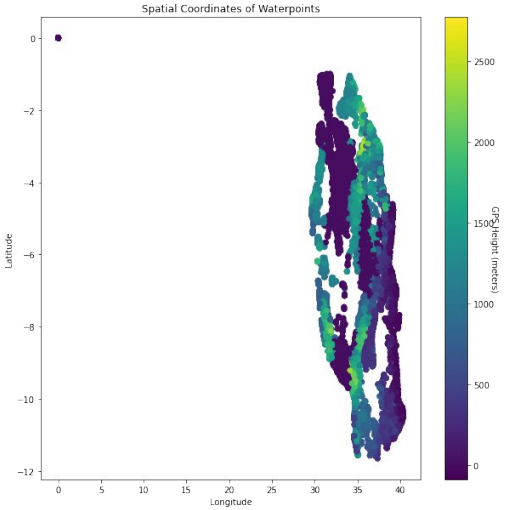
Date was provided by the Pump it Up: Data Mining the Water Table competition on Driven Data:

- https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/

Goal

Produce a model to classify the status of waterpoints in Tanzania with the highest possible accuracy score. The possible classes are:

- Functional
- Functional Needs Repair
- Non-Functional



Predictors: Geospatial

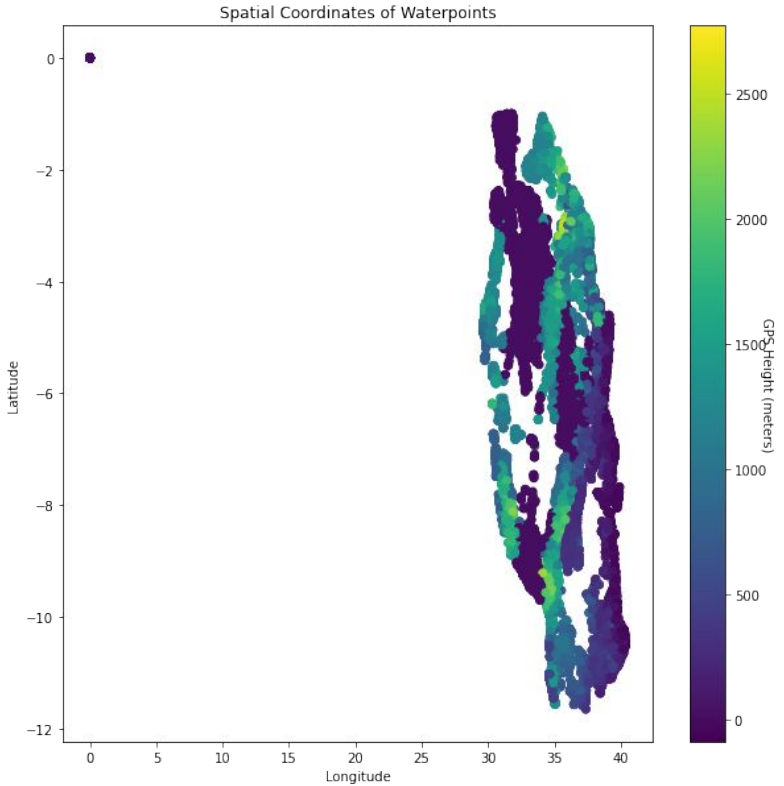
We have Geospatial data in the following features:

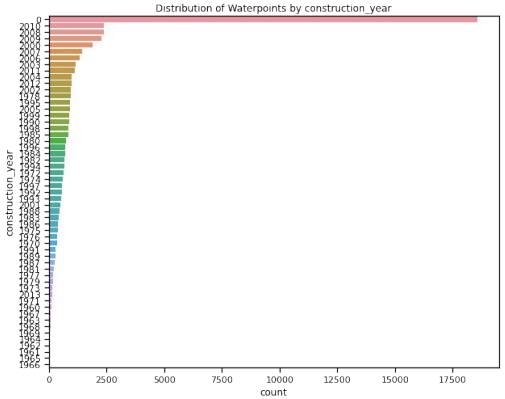
- Longitude
- Latitude
- GPS Height

About 3% of data is missing geospatial coordinates, encoded with zeros.

GPS Height feature can be negative. Possibly incorporates well depth.

Predictors: Geospatial





Predictors: Installation

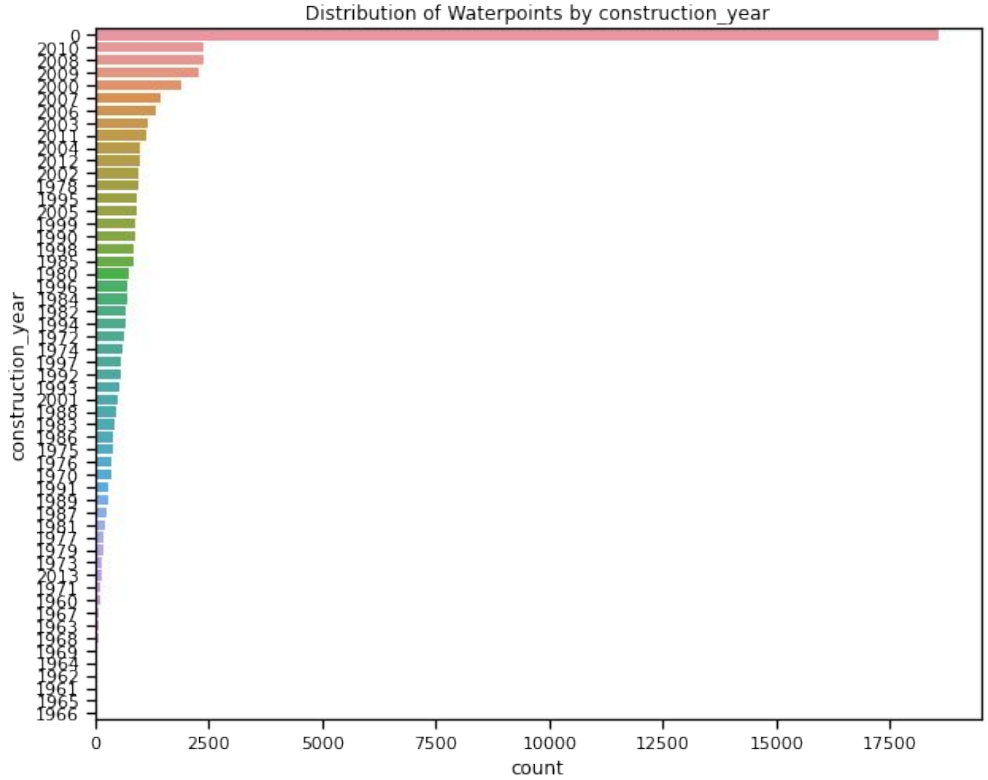
Data about the installation of water points in the features:

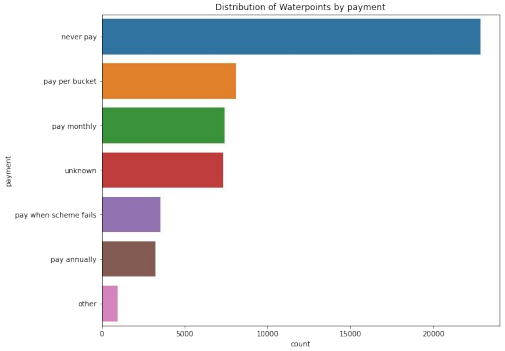
- Construction Year
- Installer
- Funder

Both Installer and Funder have far too many classes to inspect visually.

Construction Year has missing values encoded with zeros.

Predictors: Installation





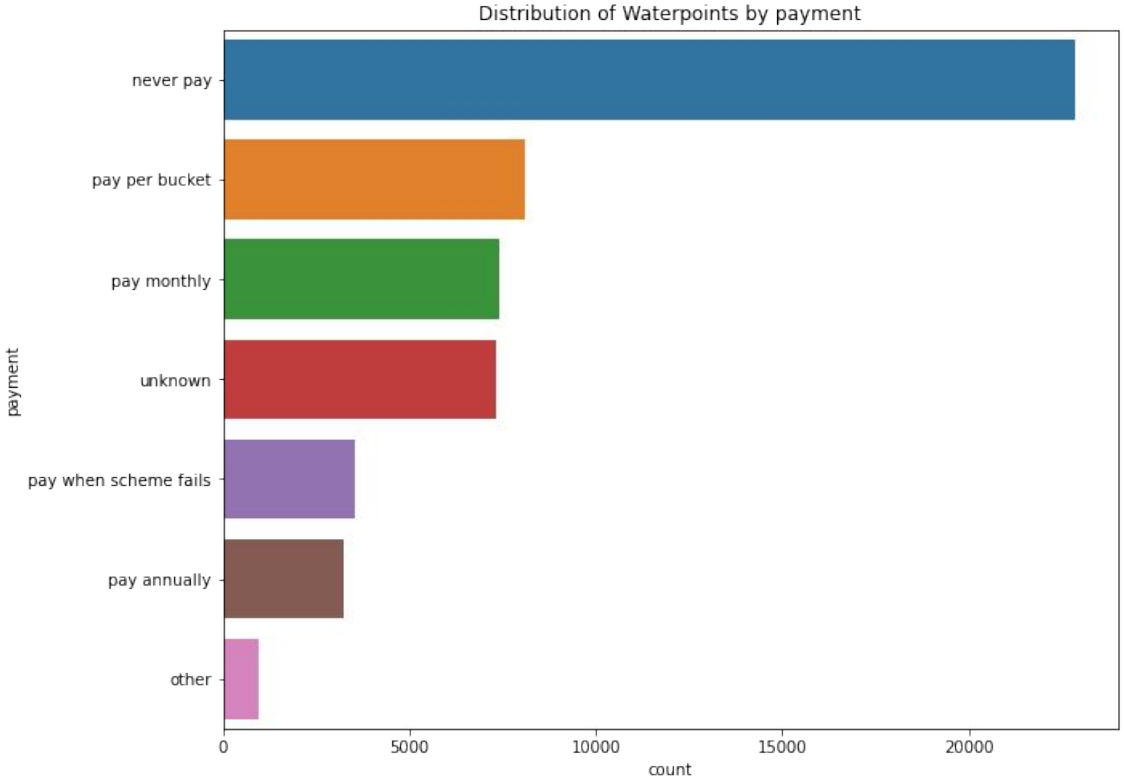
Predictors: Management

Waterpoint management data in the following features:

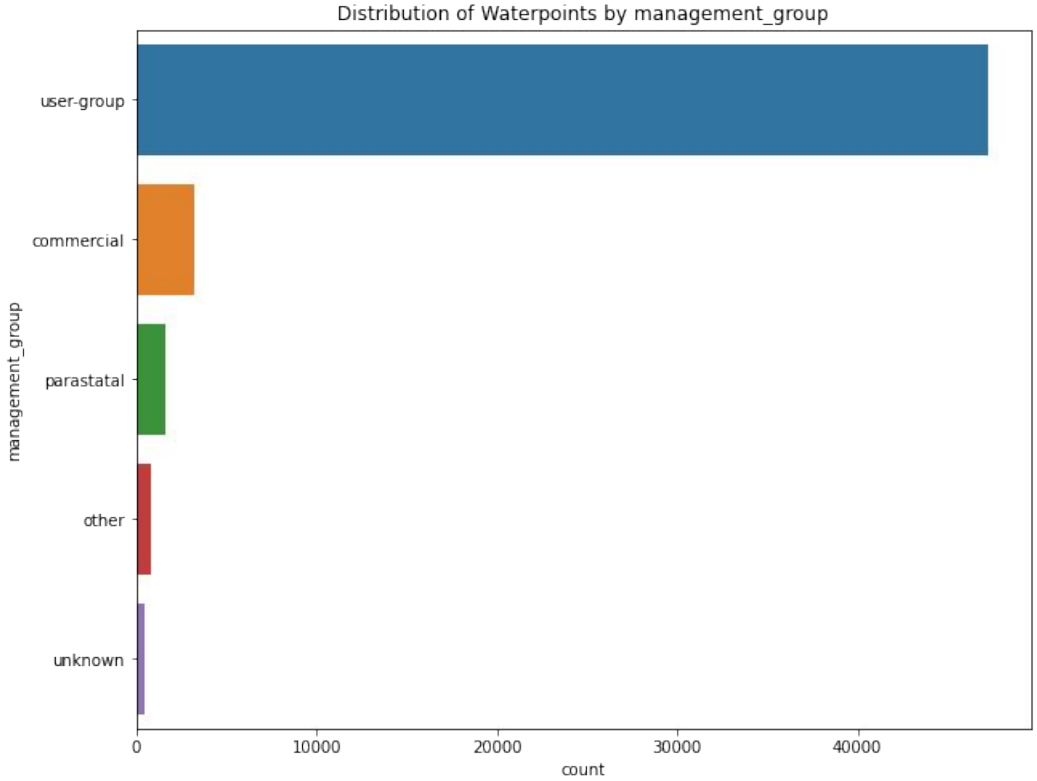
- Scheme Management
- Scheme Name
- Management
- Management Group
- Payment
- Payment Type
- Permit

Scheme Name has too many classes to inspect visually.

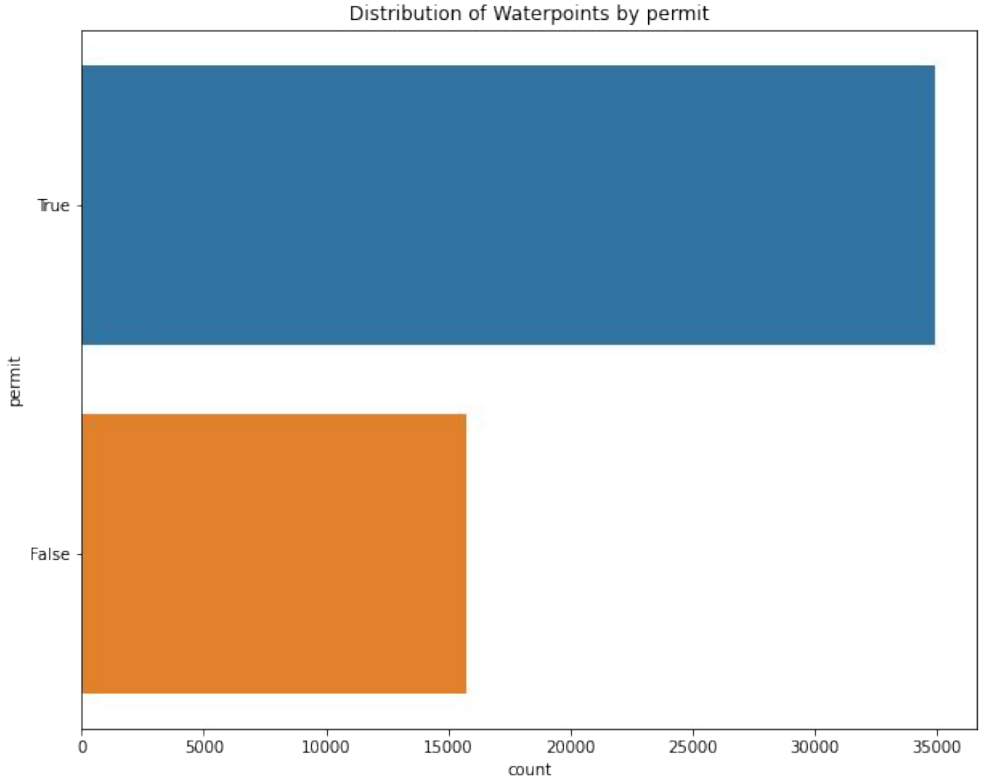
Predictors: Payment

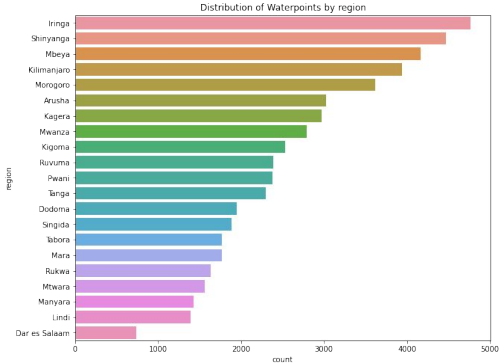


Predictors: Management Group



Predictors: Permit





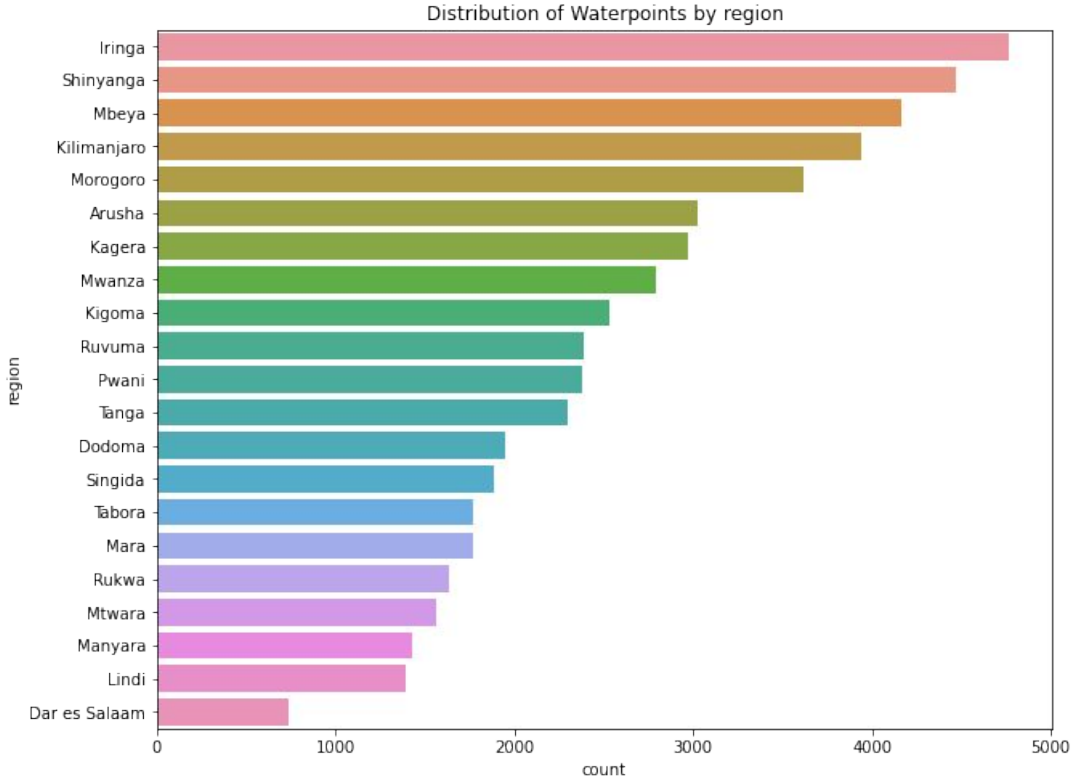
Predictors: Regional

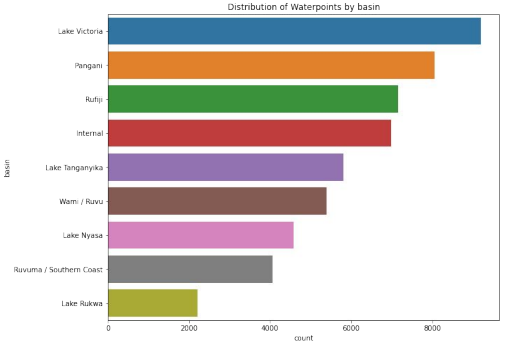
Water points are located all over Tanzania. We have regional data at four levels:

- Region
- District
- Ward
- Sub-Village

All but the top level Region data have too many classes to inspect visually.

Predictors: Region



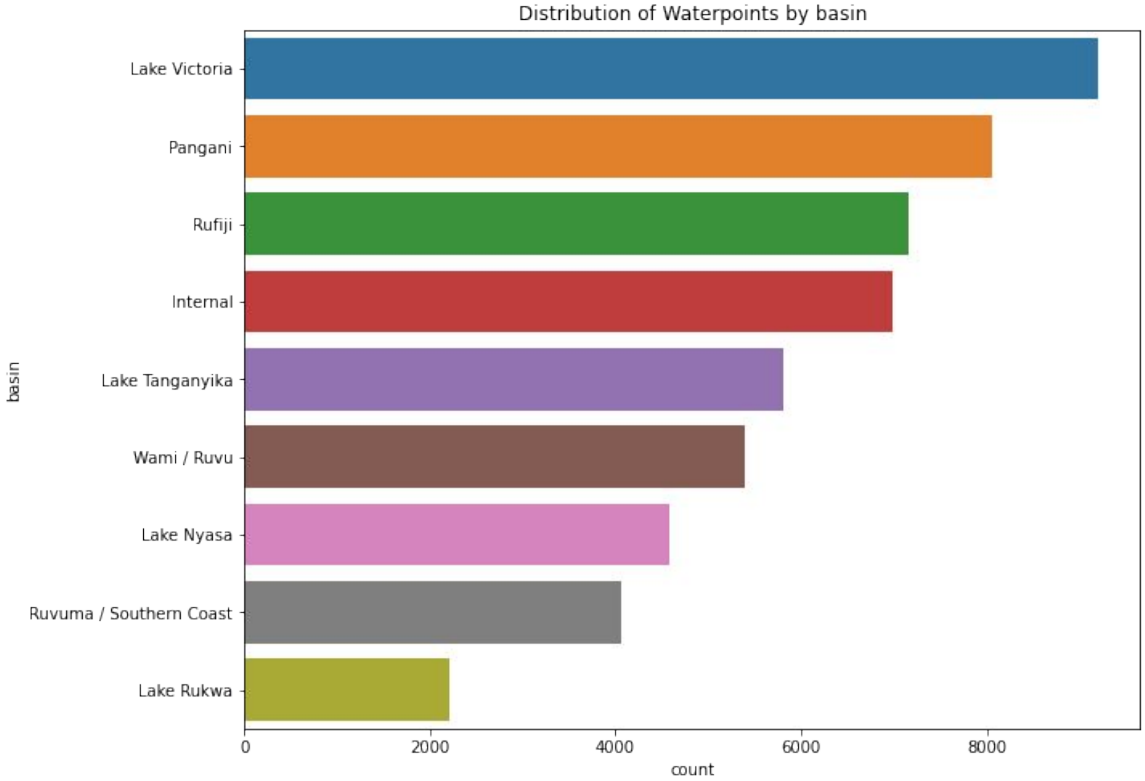


Predictors: Water

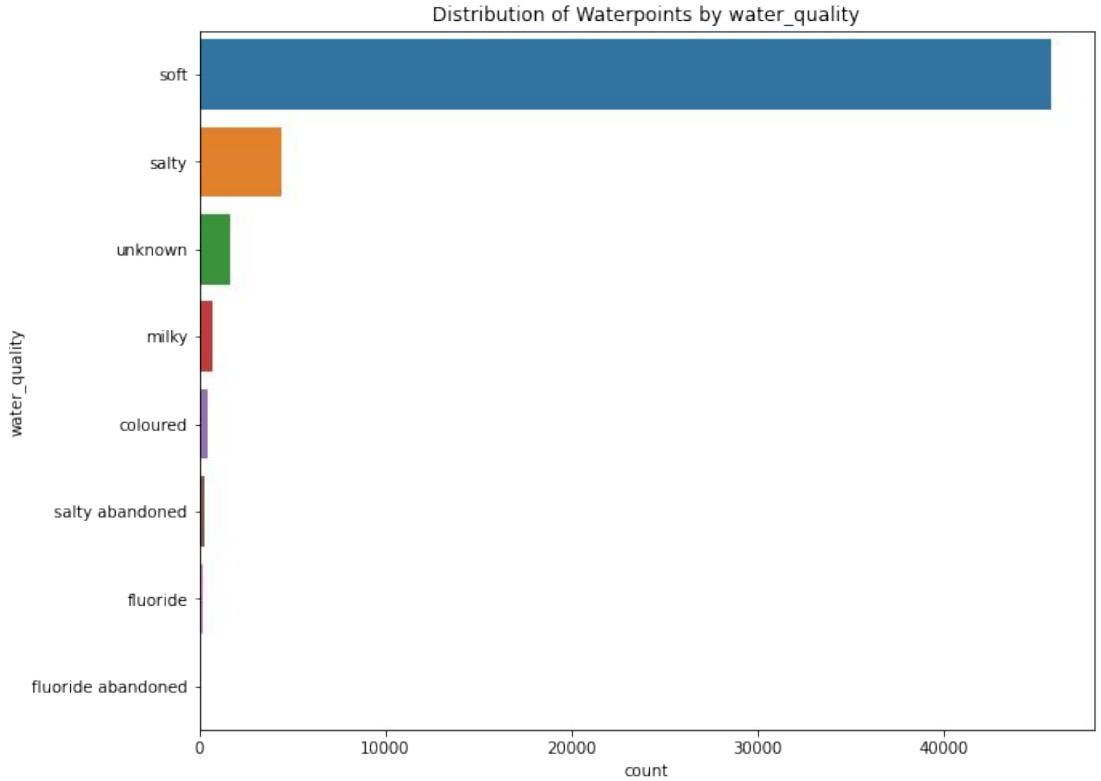
Information about the water accessed by the waterpoint:

- Basin
- Water quality
- Quality Group
- Quantity
- Source
- Source Type
- Source Class

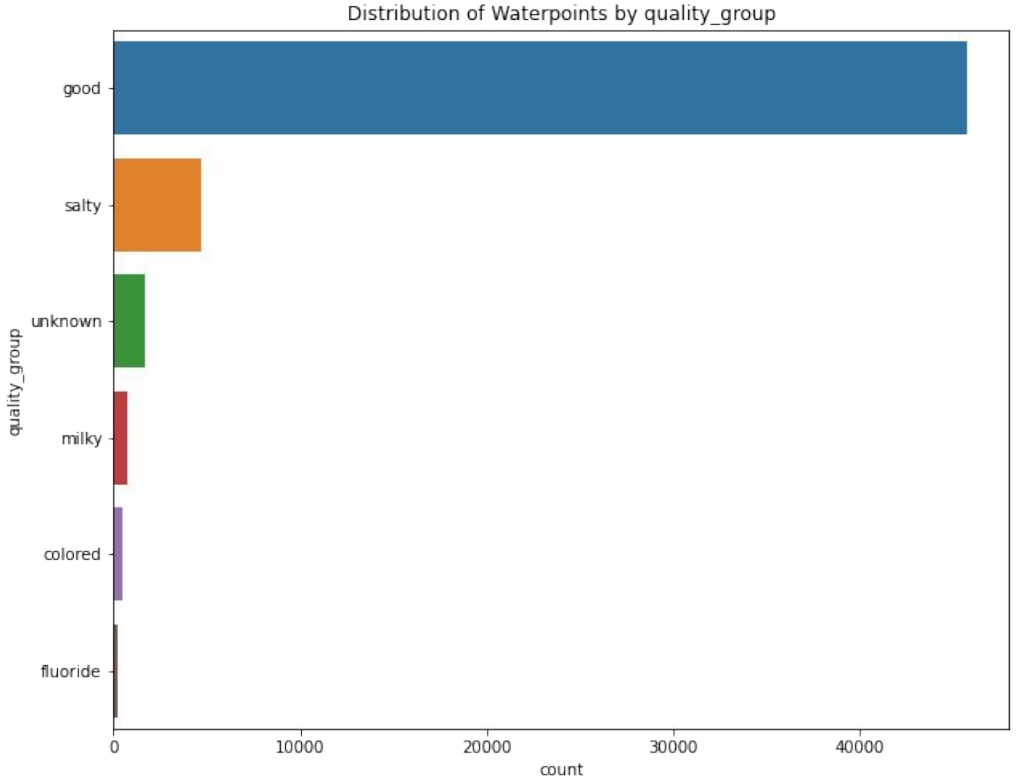
Predictors: Basin



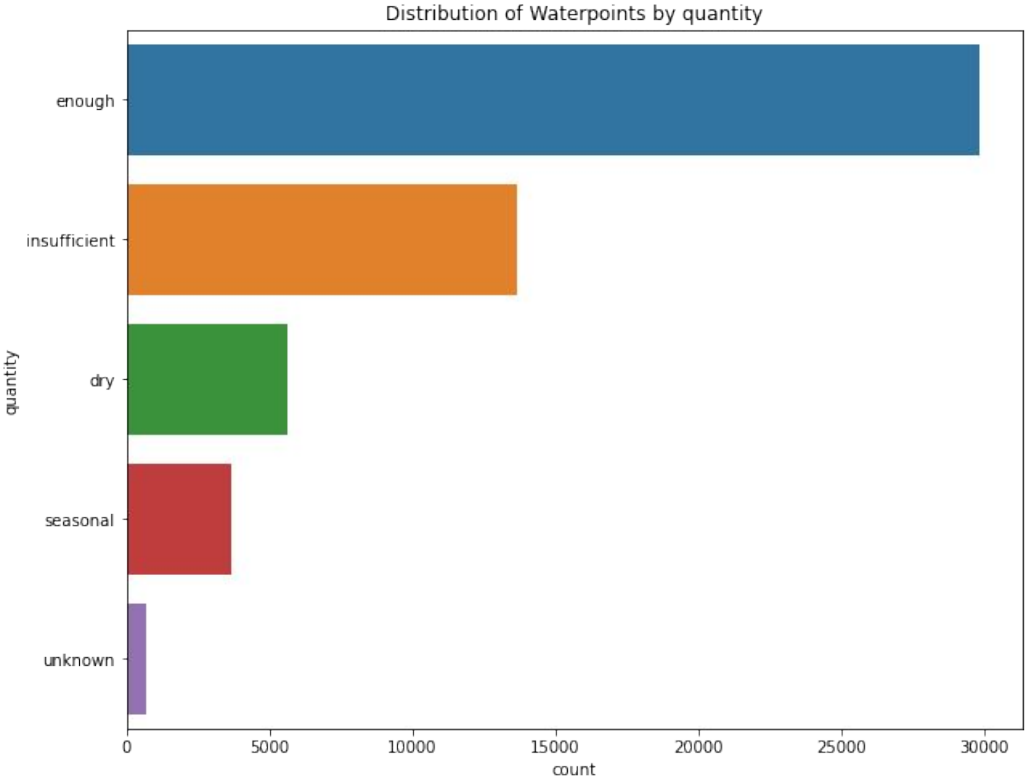
Predictors: Water Quality



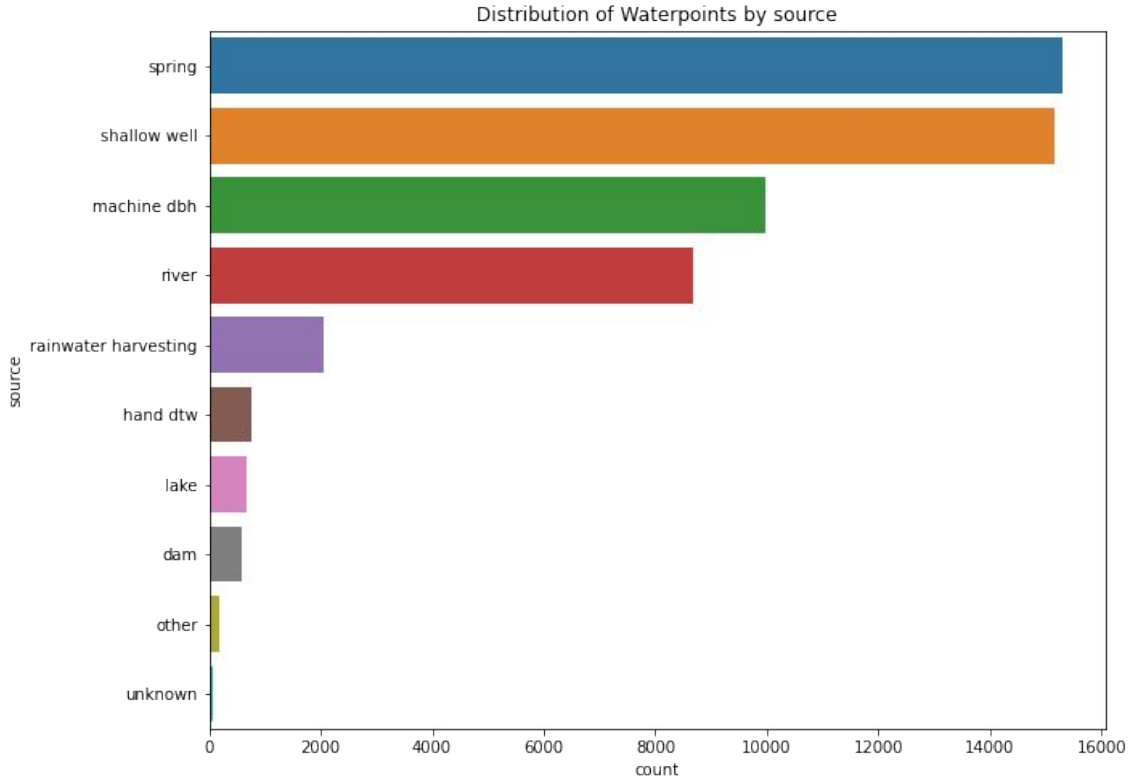
Predictors: Quality Group



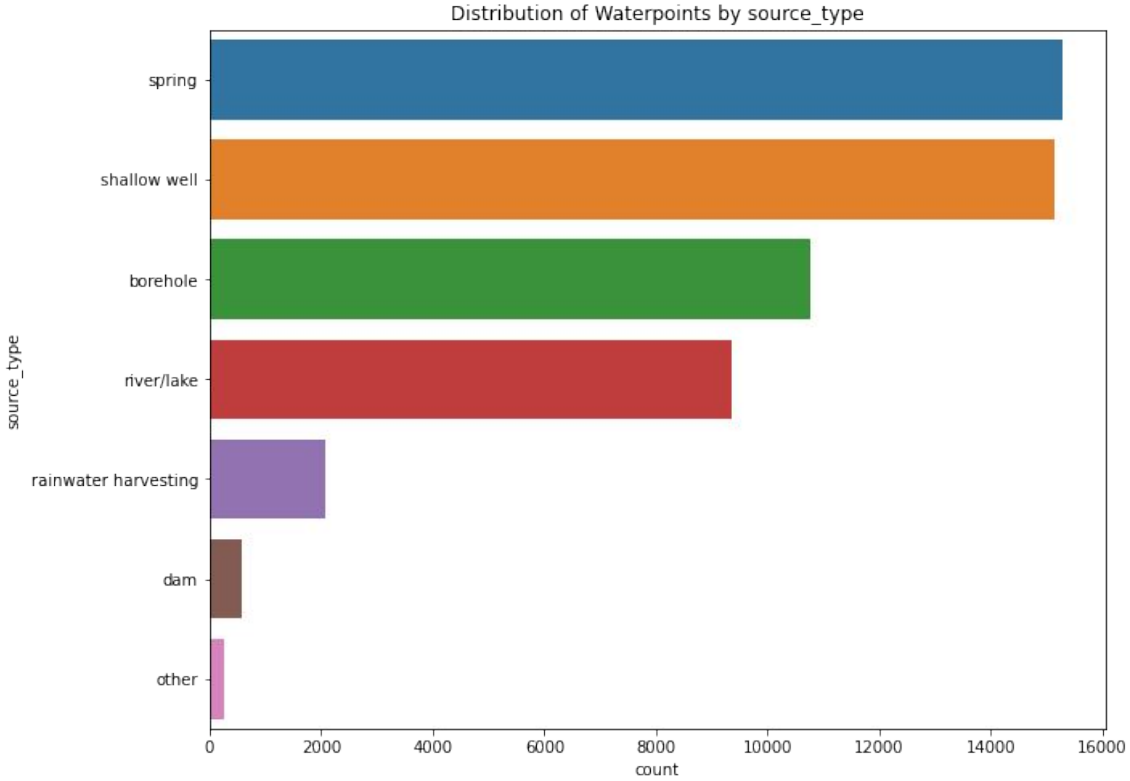
Predictors: Quantity



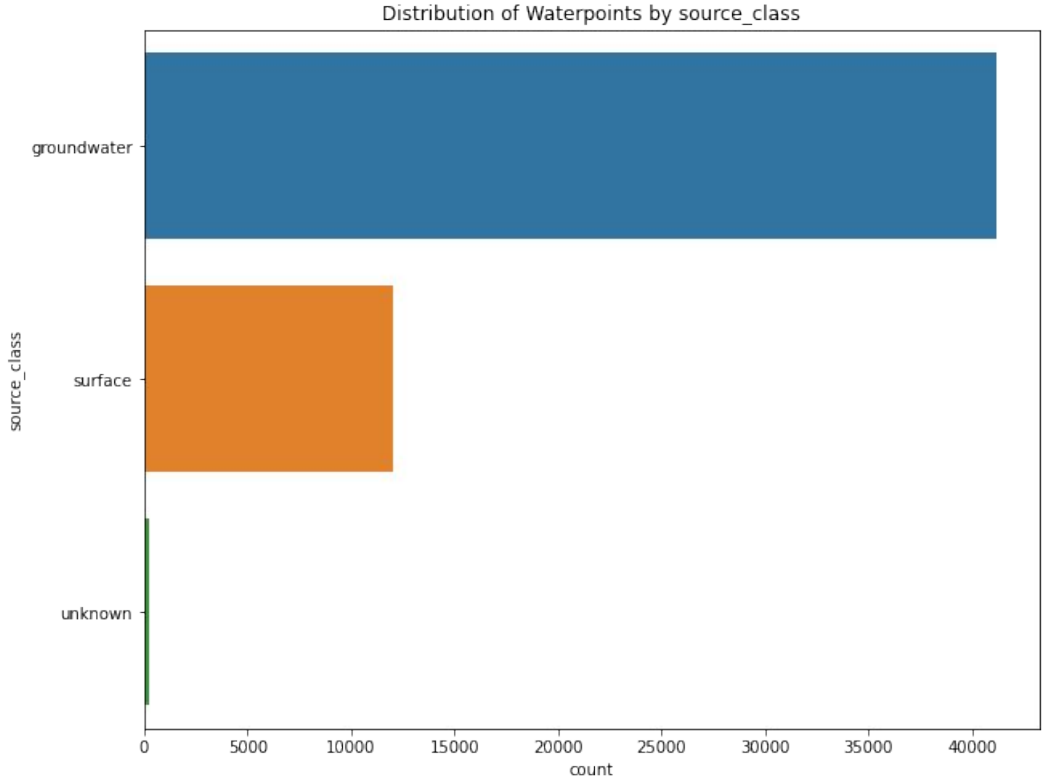
Predictors: Source

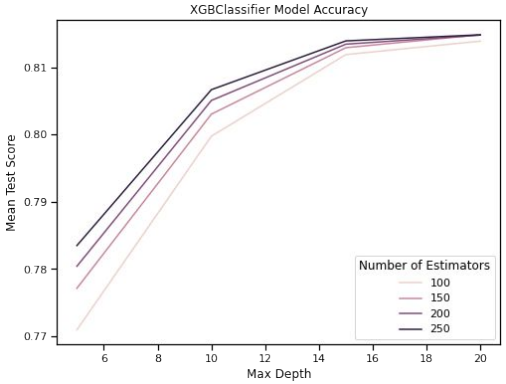


Predictors: Source Type



Predictors: Source Class



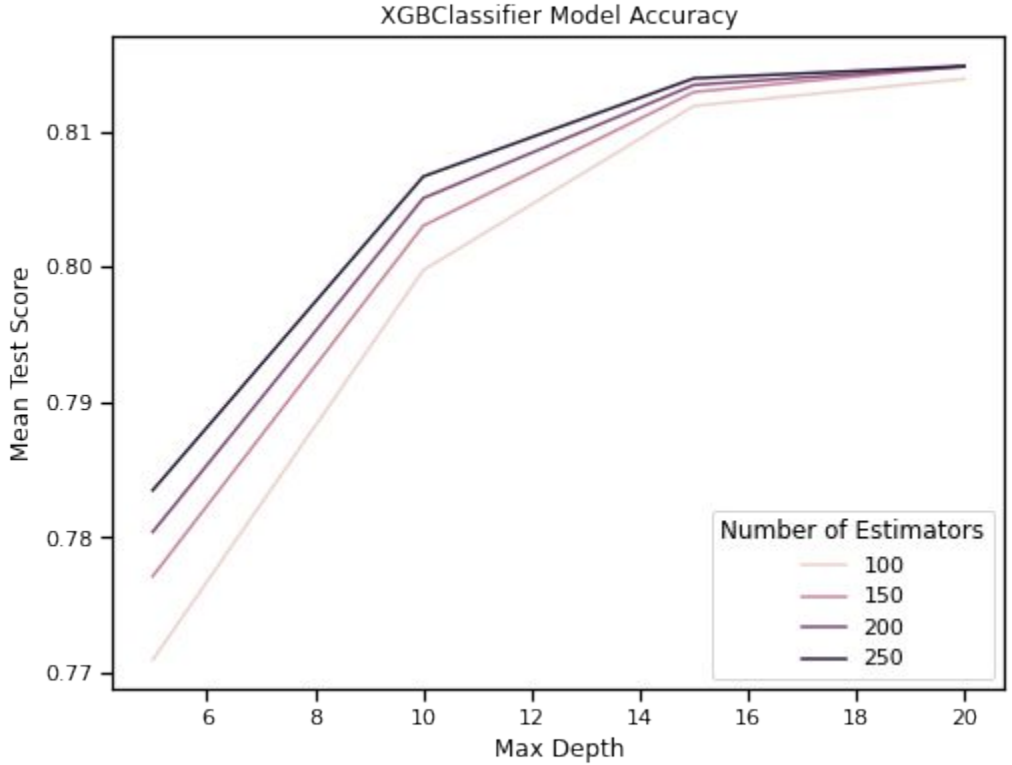


The model

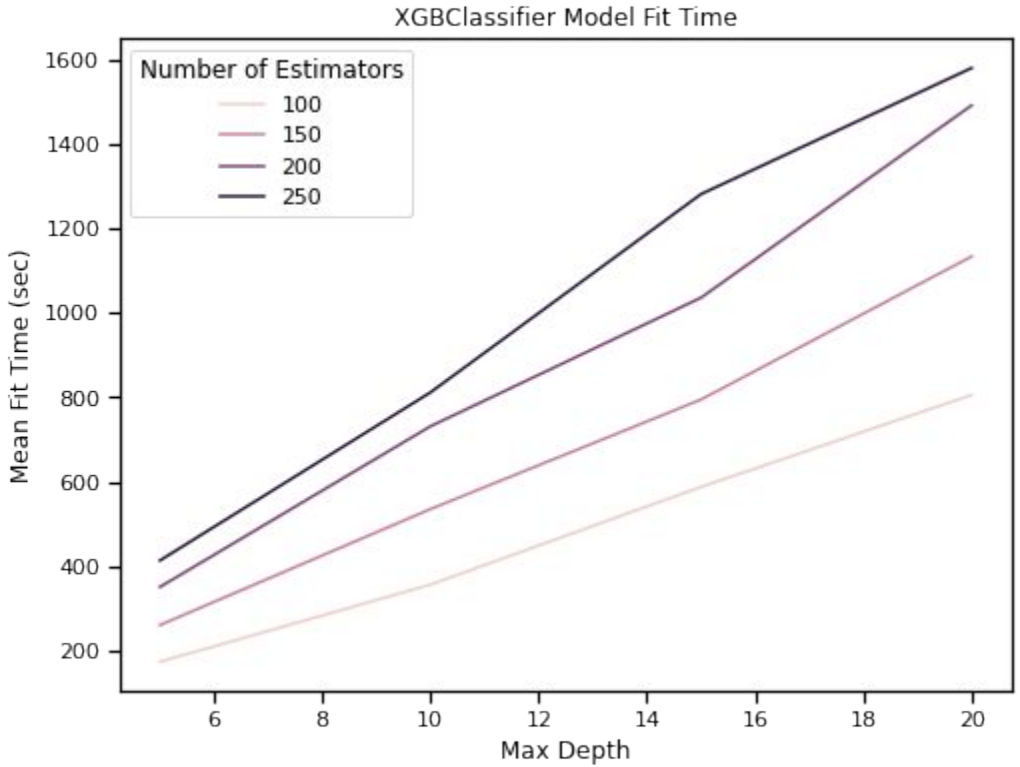
The final model was a XGBoosted random forest classifier with a maximum tree depth of 20 splits and 200 learners.

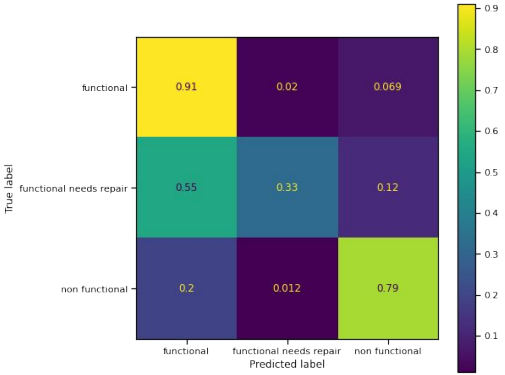
This model was selected using a cross validated gridsearch.

Model Evaluation: Accuracy



Model Evaluation: Fit Time

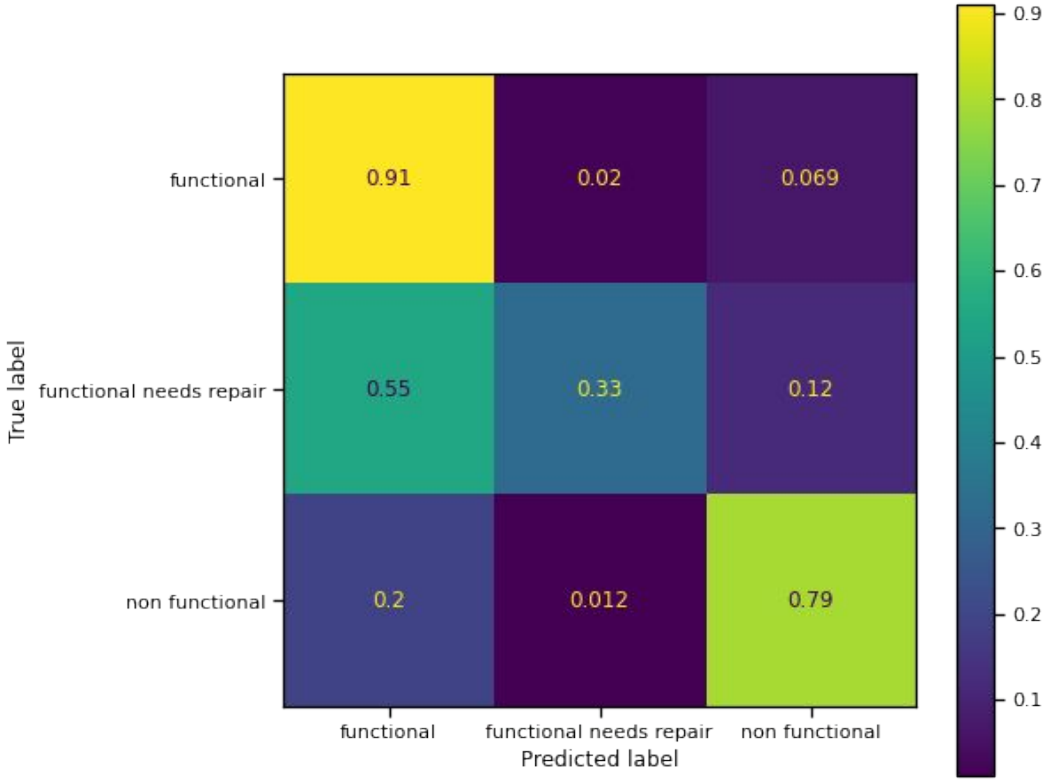




Conclusions

- Accuracy is worst on the 'functional needs repair' group.
- This class is under-represented in the data.
- Most likely, ambiguously defined in comparison to the other classifications.

Confusion Matrix



Future Work

The two most promising directions for further work:

- Integrating re-sampling into the pipeline to improve accuracy on the 'functional needs repair' class.
- Implementing hierarchical models or stacked models.

Thank You

https://github.com/sethchart/Pump-it-Up-Data-Mining-the-Water-Table