Module 3
Final Project

Will Dougherty
Flatiron School - Data Science - Online Self-Paced

Predicting Water Pump Functionality In Tanzania

Dataset: 74,250 water pumps, with 38 features

These provide water for >60 million residents,
spread over nearly a million square kilometers

Business problem:

For each water pump, there are 3 target classes:

- Functional
- Non-functional
- Functional, needs repair

This project will:

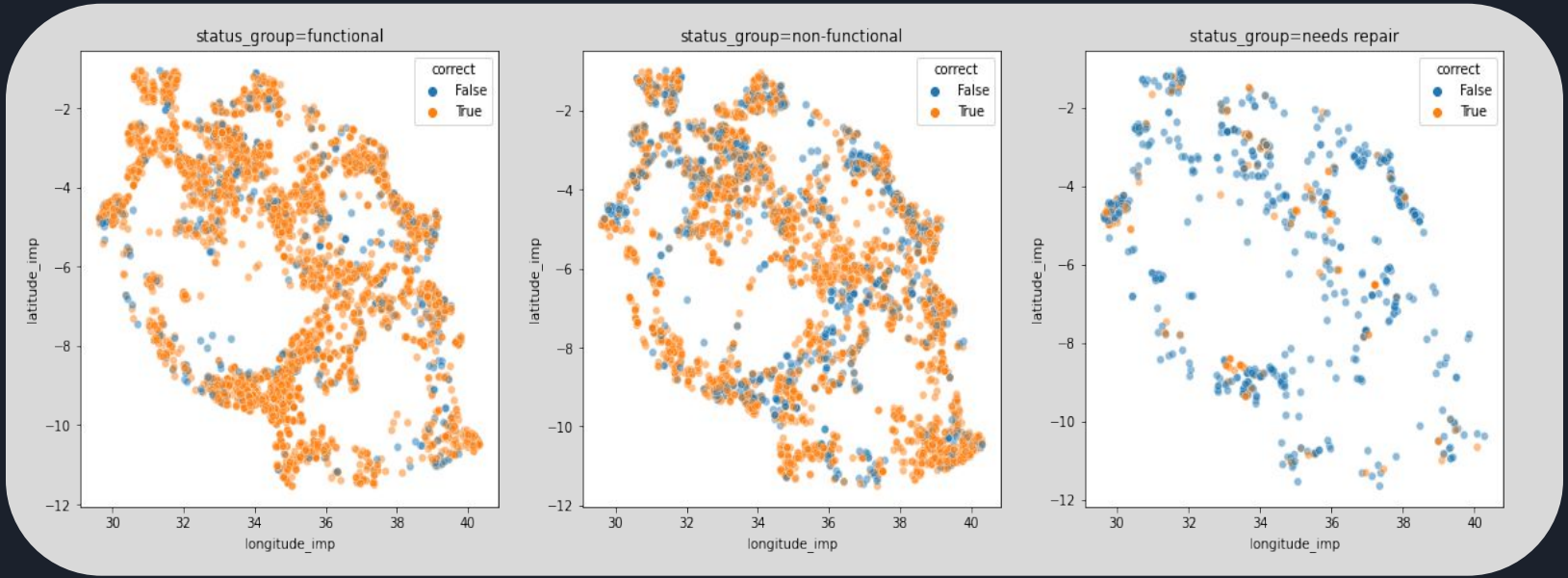
- Predict the class of a pump based on given features
- Provide insights and recommendations for action and future work, based on data analysis and results of the modelling process

Model and Results

Ensemble classification model

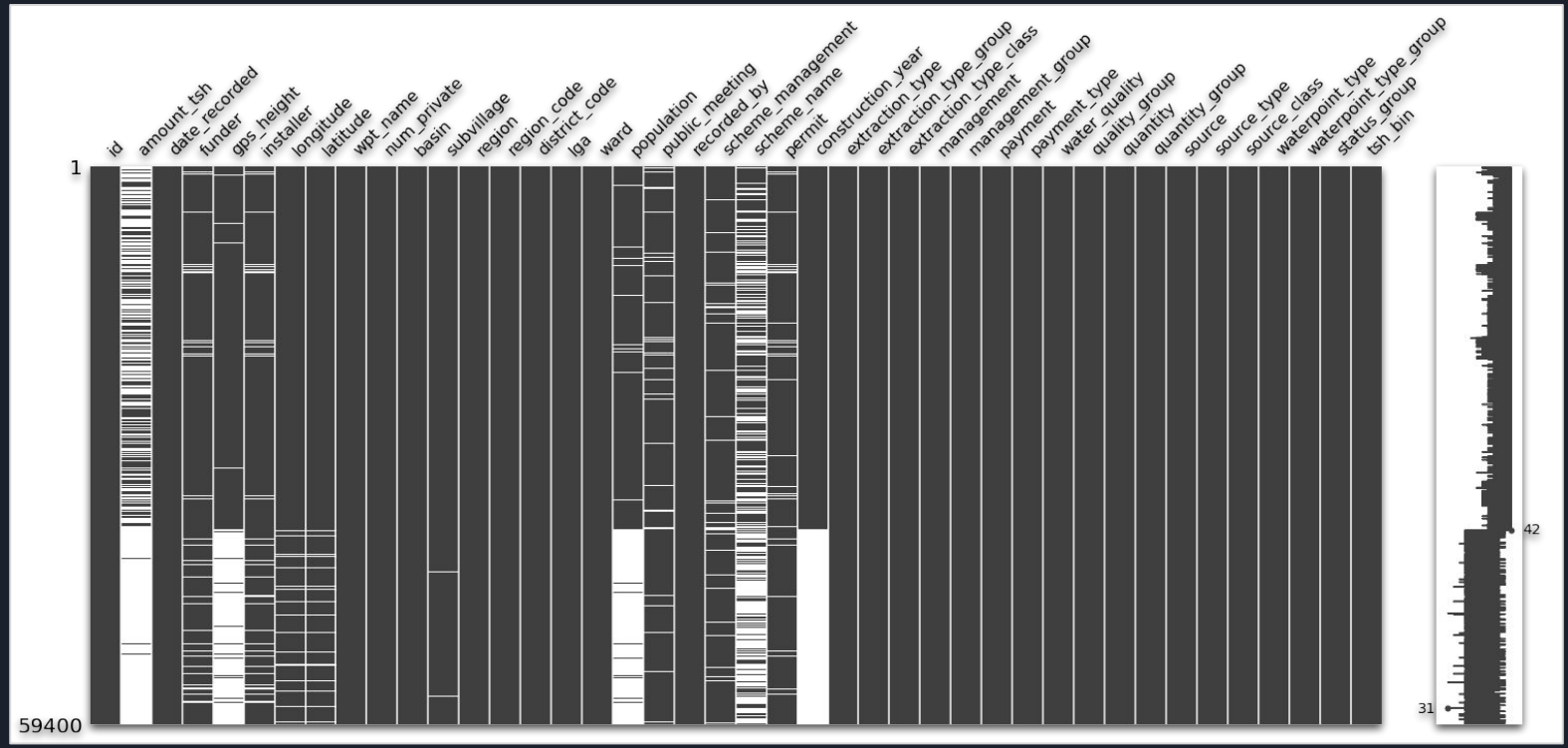
- Random forest, Bagging, XGBoost in a Voting Classifier

Accuracy: **>82%**



Main Data Problems

1: Missing-ness - ‘construction_year’

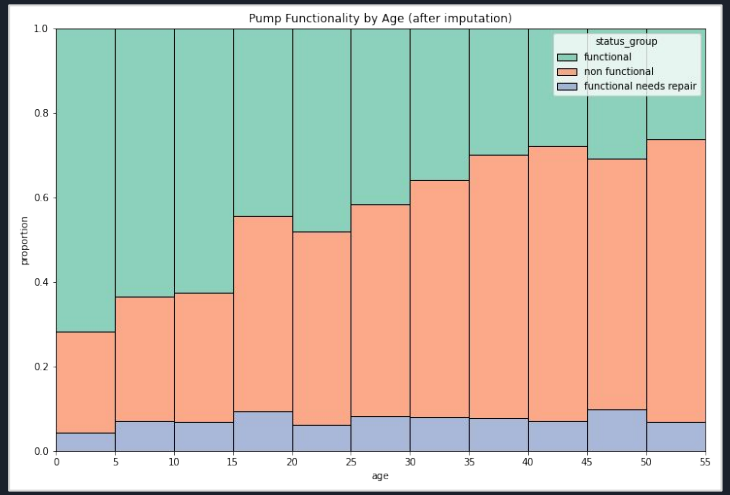


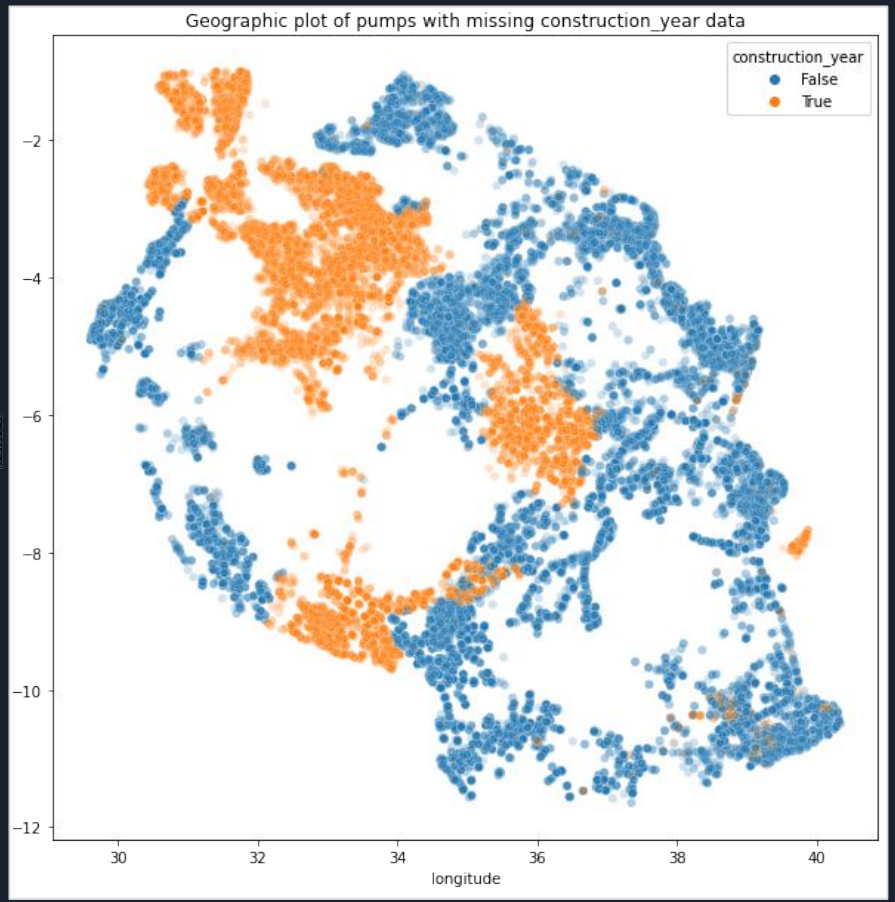
In pumps missing ‘construction_year’:

- Lack ‘amount_tsh’, ‘funder’, ‘installer’, ‘gps_height’, ‘longitude’/’latitude’, and ‘population’ in greater proportion

- Right: They are heavily clustered geographically

- Below: ‘age’ feature correlates strongly with functionality





2: Non-standardization

‘funder’ and ‘installer’ features have thousands of unique values

Many appear to be misspelled or have inconsistent spaces/punctuation

Right: highest-similarity installer names

```
{'danida || danid': 0.909,
 'community || communit': 0.941,
 'gover || govern': 0.909,
 'tasaf || tassaf': 0.909,
 'fini water || fin water': 0.947,
 'oxfam || oxfarm': 0.909,
 'kiliwater || kili water': 0.947,
 'kiliwater || kilwater': 0.941,
 'rc church || rc churc': 0.941,
 'water aid || wateraid': 0.941,
 'consulting engineer || consuting engineer': 0.973,
 'muwsa || muwasa': 0.909,
 'finwater || fin water': 0.941,
 'villa || villag': 0.909,
 'fin water || finn water': 0.947,
 'adra/community || adra /community': 0.966,
 'adra/community || adra/ community': 0.966,
 'adra /community || adra/ community': 0.933,
 'local  technician || local technician': 0.97,
 'water aid /sema || water aid/sema': 0.966,
 'jandu plumber co || jandu plumber  co': 0.97,
 'muwasa || mtuwasa': 0.923,
 'tuwasa || mtuwasa': 0.923}
```

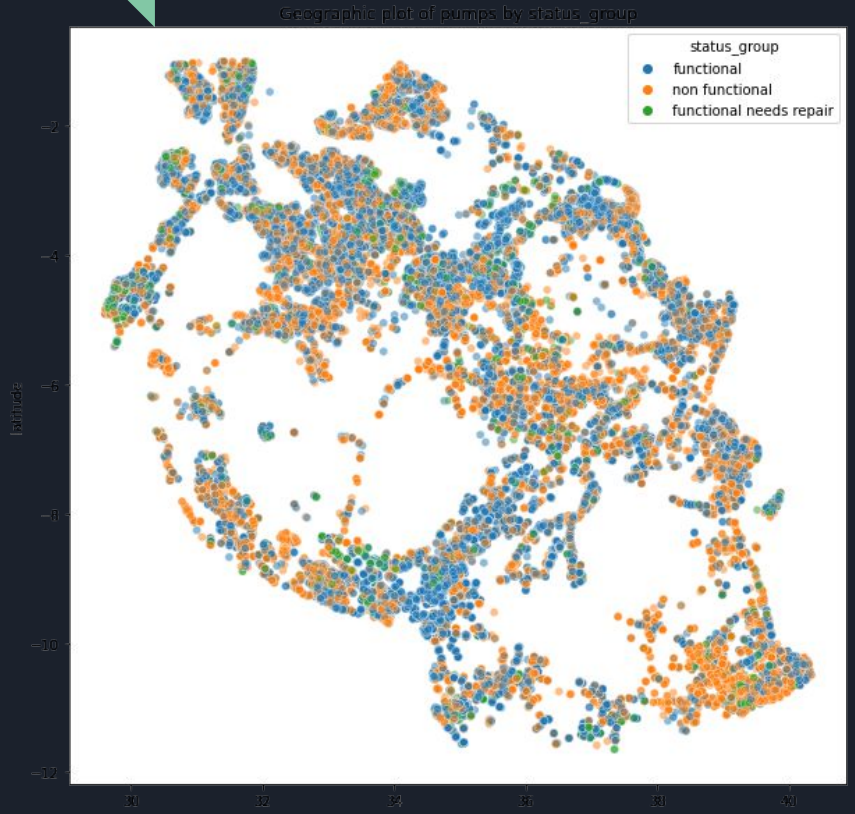
Recommendations

- To the extent that it's possible, gathering/updating of the dataset should focus on those areas lacking construction_year data, to maximize the data gain, as the lack of many other features correlate with the lack of construction_year

- A comprehensive audit of installer and funder data
    - Consulting local experts to determine correct names, especially those in native languages

- Use model for regional prediction vs individual pumps

Regional Prediction



Functionality is clustered throughout the region

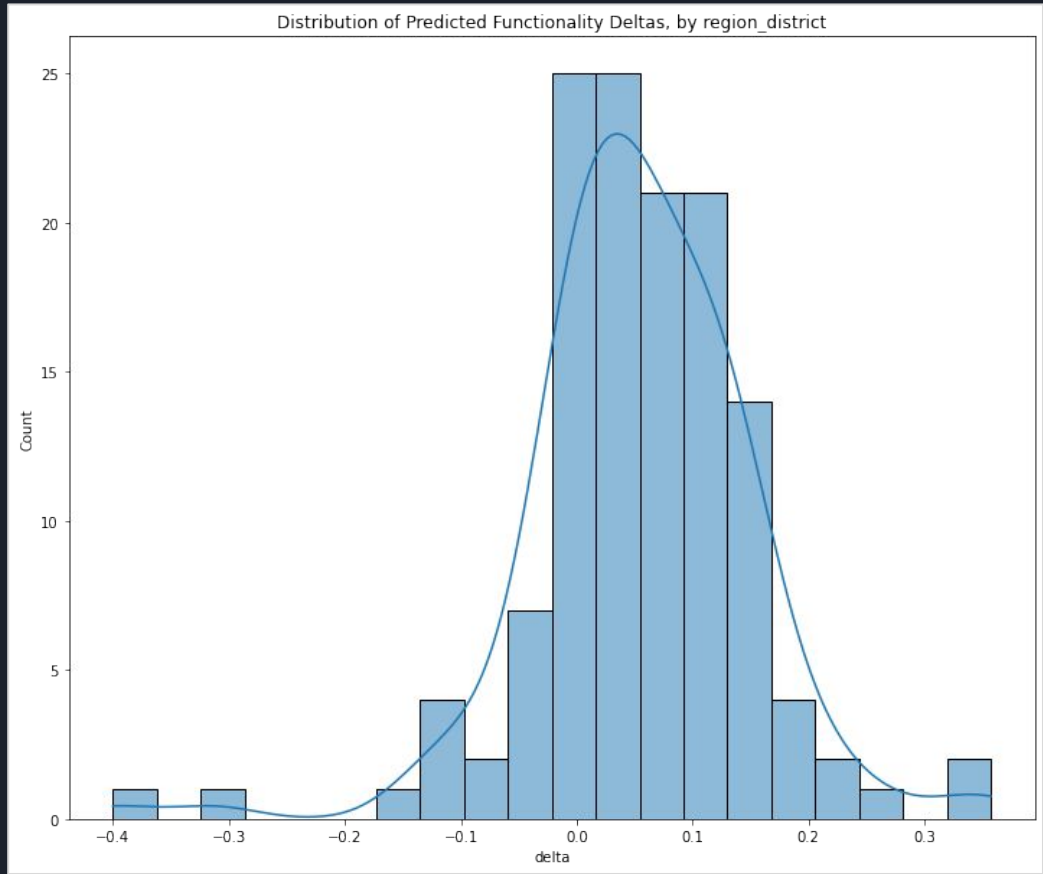
Knowing about individual pumps is less useful than knowing about areas/regions

Tentative Results

Using an engineered combination feature, ‘region_district’ - with 131 unique values

The vast majority are predicted within 10-15%

The model tends to overestimate functionality



Future Work

- Implementing recommendations to improve ‘funder’, ‘installer’
- Further work to improve imputation of missing data in construction_year and its correlated features
- Fine-tuning and further implementation of Regional Prediction
- Investigate re-classification - combine ‘needs repair’ with ‘non-functional’?

Thank you!