Module 3 Final Project

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

Predicting Water Pump Functionality In Tanzania

Dataset:

- 74,250 water pumps
- 38 features

Tanzania:

- 60 million residents
- Over 1 million km^2

UGANDA
Lake Victoria
Bukoba
Buoen
Musoma
RWANDA
KAGERA
MARA
KENYA
UNITED REP. OF TANZANIA
Ukerewe I.
Ukara I.
Rubondo I.
Maisome I.
Mwanza
Ngara
Biharamulo Game Res.
SERENGETI NAT. PARK
Lake Natron
MT. KILIMANJARO NAT. PARK
Kilimanjaro 5895 m
Geita
MWANZA
Ngorongoro Conservation Area
Maswa Game Res.
Arusha
Moshi
BURUNDI
Kibondo
SHINYANGA
Kahama
Shinyanga
L. Eyasi
KILIMANJARO
Same
Mkomazi Game Res.
Nzega
Lake Manyara
Babati
TARANGIRE NAT. PARK
ARUSHA
Masai Steppe
KIGOMA
Kasulu
Kigoma
Ujiji
Iwembere Steppe
Singida
Kondoa
Mkomazi
Tanga
INDIAN OCEAN
Wete
Pemba
Tobora
Kaliua
SINGIDA
Korogwe
Handeni
TANGA
Pangani
ZANZIBAR
Zanzibar
Manyoni
Ugalla River Game Res.
TABORA
Dodoma
Mpwapwa
Sadani
Lake Tanganyika
Mpanda
RUKWA
KATAVI N.P.
DODOMA
Morogoro
Kibaha
DAR ES SALAAM
Dar es Salaam
Rungwa Game Reserve
RUAHA NAT. PARK
MIKUMI N.P.
PWANI
DEMOCRATIC REPUBLIC OF THE CONGO
Lake Rukwa
Iringa
Mafia I.
Sumbawanga
Mpui
MBEYA
IRINGA
MOROGORO
Utete
Mohoro
ZAMBIA
Chunya
Selous Game Reserve
Mbeya
Kilwa Kivinje
Tunduma
Tukuyu
Njombe
LINDI
KIPENGERE RANGE
Lindi
MALAWI
Lake Nyasa
Songea
Nachingwea
Mtwara
Masasi
RUVUMA
MTWARA
Newala
Tunduru
MOZAMBIQUE
UNITED REPUBLIC OF TANZANIA
National capital
Provincial capital
Town, village
Airport
International boundary
Provincial boundary
Main road
Secondary road
Railroad
0 100 200 300 km
0 100 200 mi
30°
32°
34°
36°
38°
2°
4°
6°
8°
10°
12°

Problem:

- Limited resources in a large country
- Up-to-date functionality data is hard to come by

Project Goals:

- Predict **Non-Functionality**
- Use model to show how and where to efficiently allocate resources

Methodology

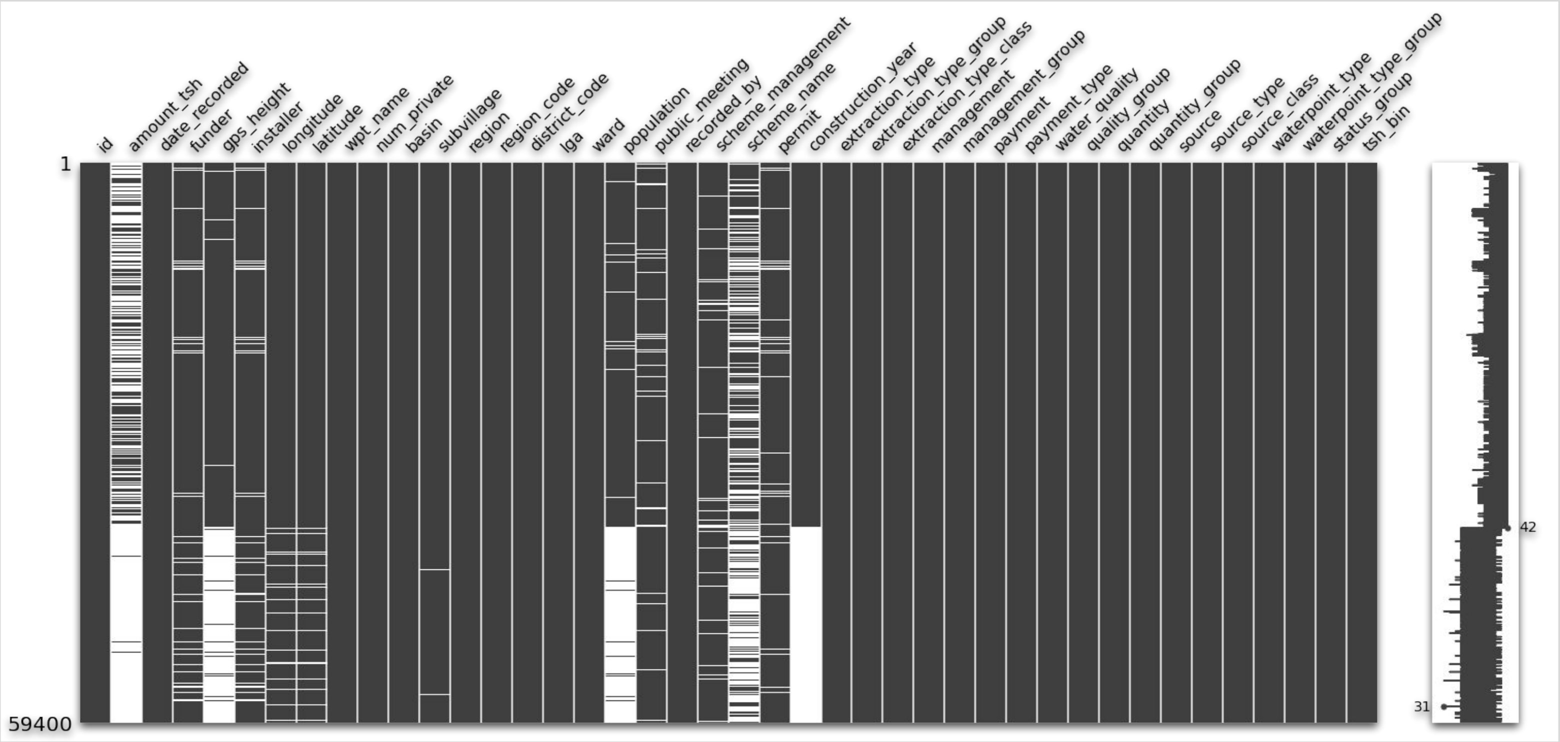
Focus on **Recall**

- Maximize % of non-functional pumps correctly identified
- Minimize false positives
    - mis-identifying *non-functional* pumps as *functional*

Build a **binary classification** model **-** ensemble of decision tree/random forest/etc

Main Data Problems

1: Missing-ness - ‘construction_year’

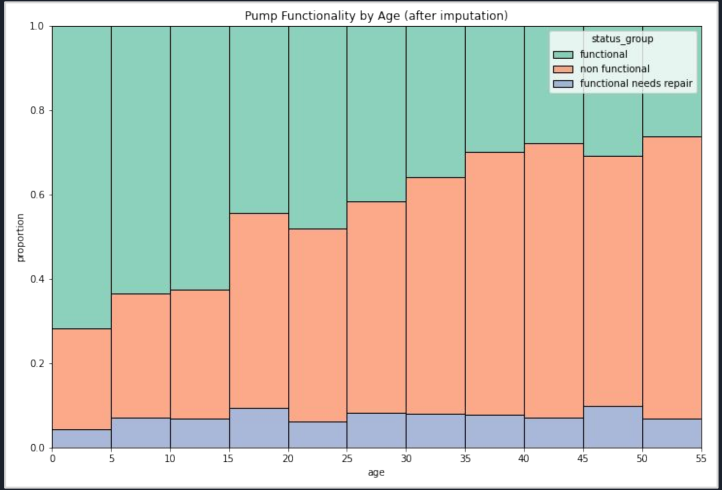


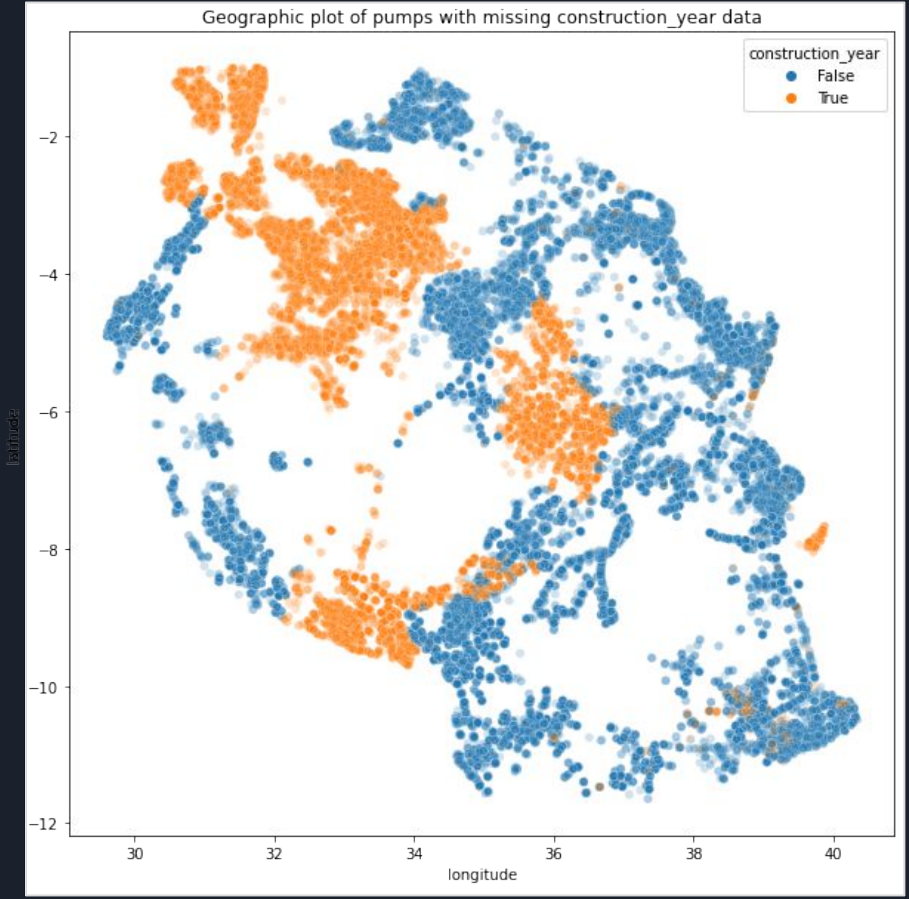
In pumps missing ‘construction_year’:

- 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



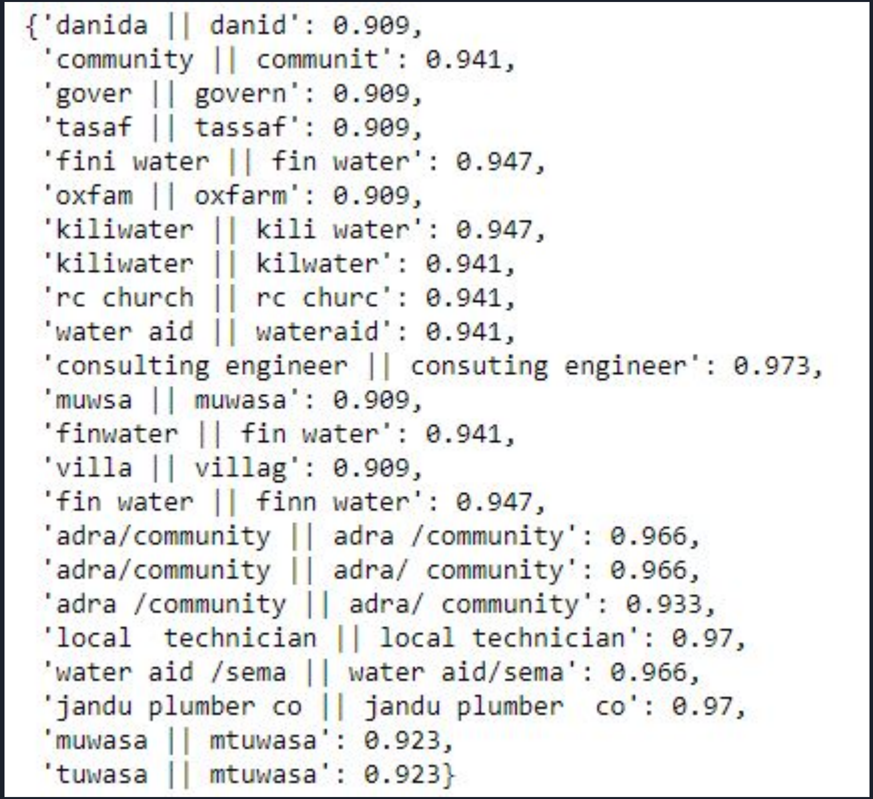


2: Non-standardization

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

Many appear to be misspelled or have inconsistent spaces/punctuation

Right: highest-similarity installer names

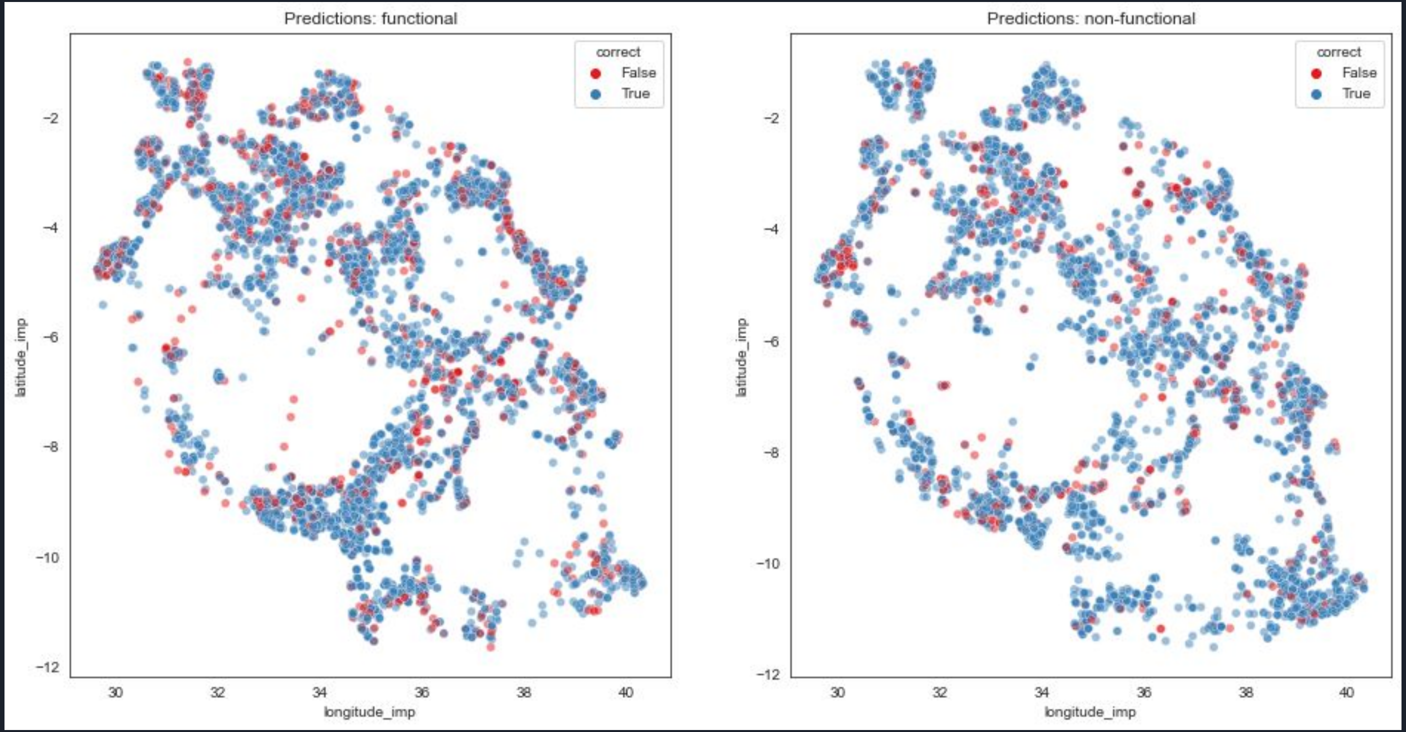


```
{'danida || danid': 0.909,
 'community || communit': 0.941,
 'gover || govern': 0.909,
 'tasaf || tassaf': 0.909,
 'fini water || fin water': 0.947,
 'oxfam || oxfarm': 0.909,
 'kiliwater || kili water': 0.947,
 'kiliwater || kilwater': 0.941,
 'rc church || rc churc': 0.941,
 'water aid || wateraid': 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}
```

The Model

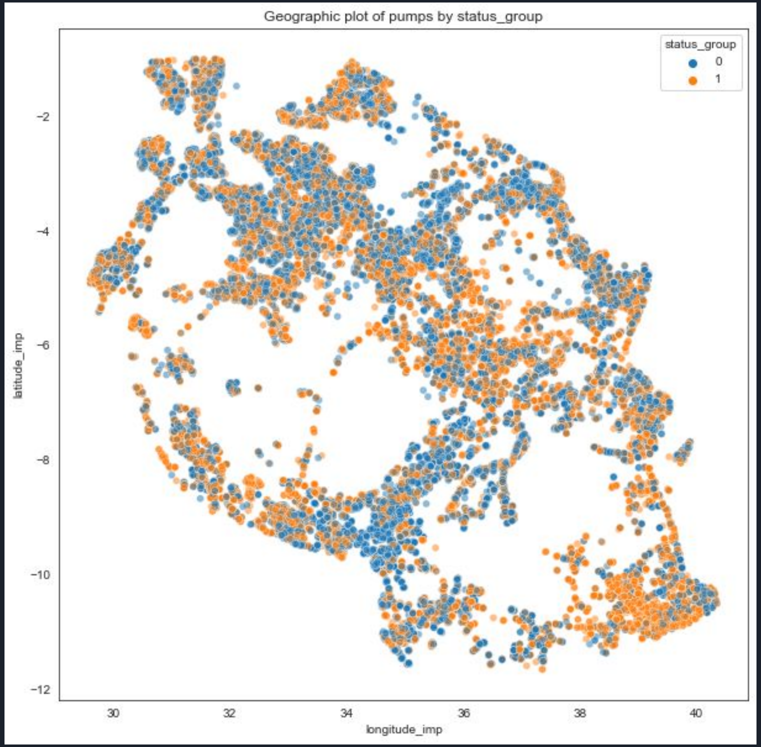
Ensemble classification model:

- Decision tree, Random forest, Bagging, XGBoost in a Voting Classifier



Recall: 85% of non-functional pumps are correctly identified

Regional Prediction



Functionality is clustered throughout the region

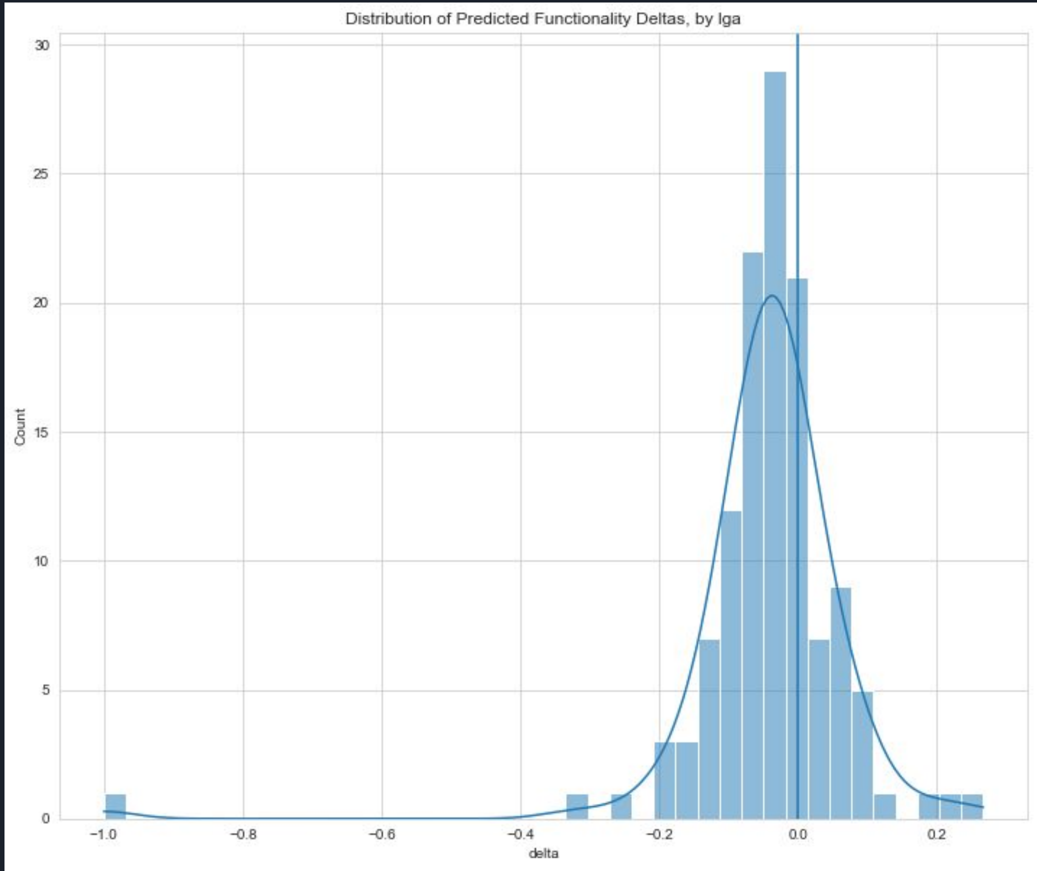
Knowing about individual pumps is less useful than knowing about areas/regions

‘Lga’ is a good middle-ground of granularity between large areas and villages

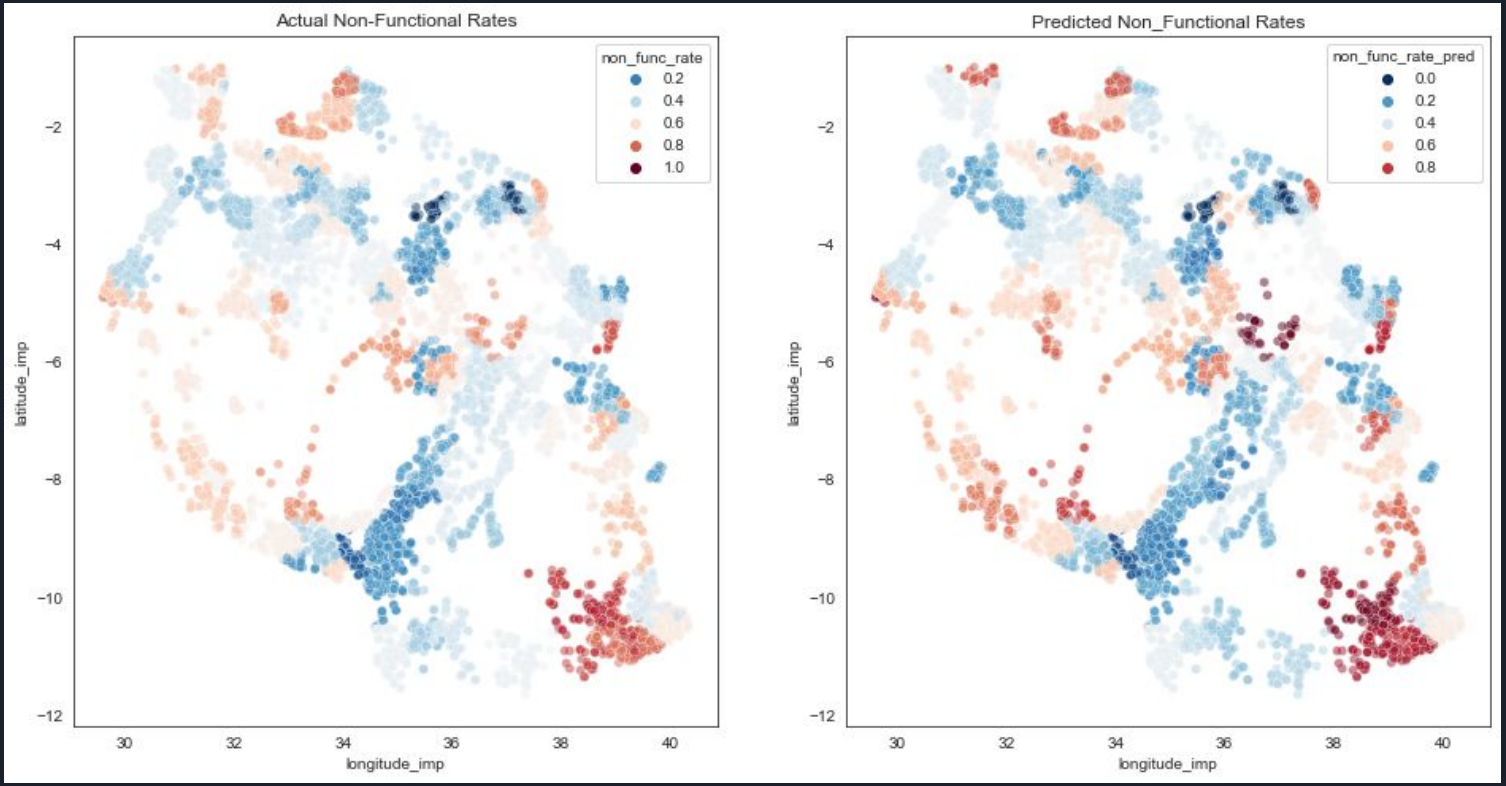
Tentative Results

The vast majority are predicted very accurately, closely packed around -0.1 - 0.0

The model tends to under-estimate rate of non-functionality compared to the actual rate



Map of non-functionality rates:
Actual vs. Predicted



Conclusions

- The model does well in the **recall** metric, by correctly predicting 85% of the non-functional pumps

- Using the model, we should focus on **regional prediction**
  - Using ‘lga’, we can accurately predict rates for areas
  - This should be used to determine areas most in need of resources

Future Work

- Improve ‘lga’ and regional prediction
  - More consistent distribution of pumps - consolidate areas with few pumps
  - Determine population within each area to make resource allocation proportional

- Determine quality of model over time
  - Predict 5 years in the future, and test those pumps in 5 years to see how it performs

Thank you!