Tanzanian Waterwell Project

Group members: Steven Kyle

Project

Tanzania has a shortage of clean water that can be provided to their citizens. The population of Tanzania is about 56 million and about 4 million citizens do not have access to clean water.

The purpose of this project is to create a model that can accurately predict if a well is functional or not. This model can then help improve maintenance operations all across Tanzania.

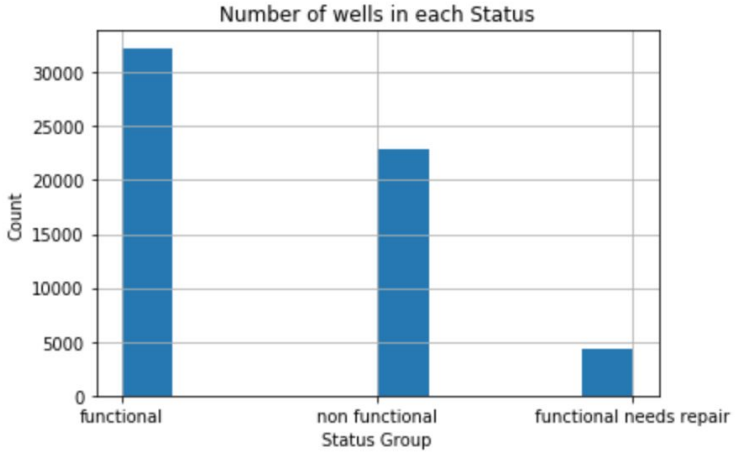
The first metric we will use to evaluate the model is having a high overall accuracy score and second having a high Recall on non-functional wells.

Water well data

Data was provided through a driven data competition by Taarifa and the Tanzanian Ministry of Water.

Snapshot of the Raw Data:
-59,400 wells
-40 features
    -6 features were continuous
    -34 features were categorical
- accounts for 10,686,653 people

functional	32259
non functional	22824
functional needs repair	4317



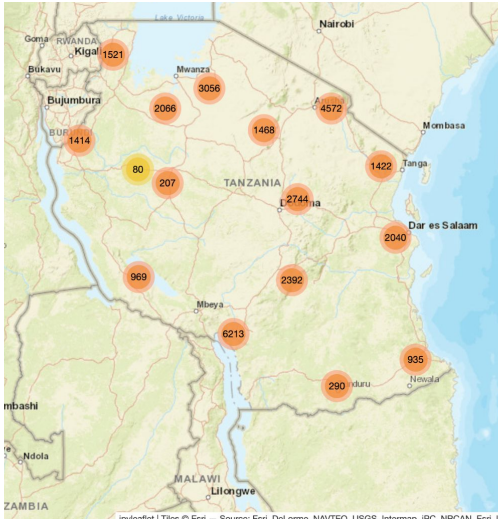
EDA

After EDA there were
- 57,588 wells
- 27 features
    - 6 continuous
    - 21 categorical
- 1,236 features after making dummy variables

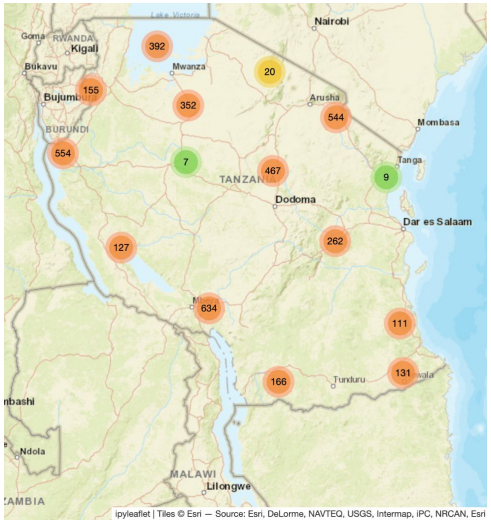
After running the final model, the model reported the top 5 most important features were
-gps height
-longitude and latitude
-population
- quantity_enough (categorical)

Longitude/Latitude

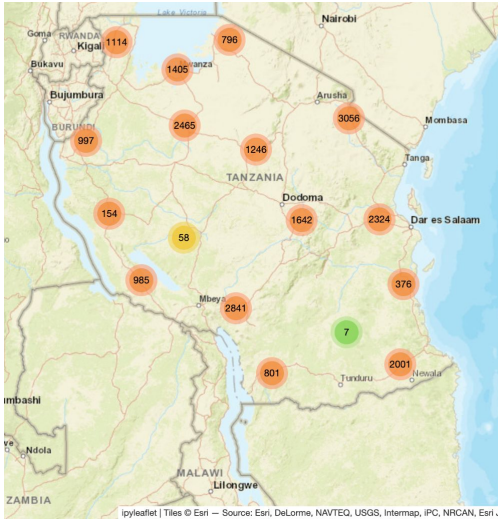
Functional



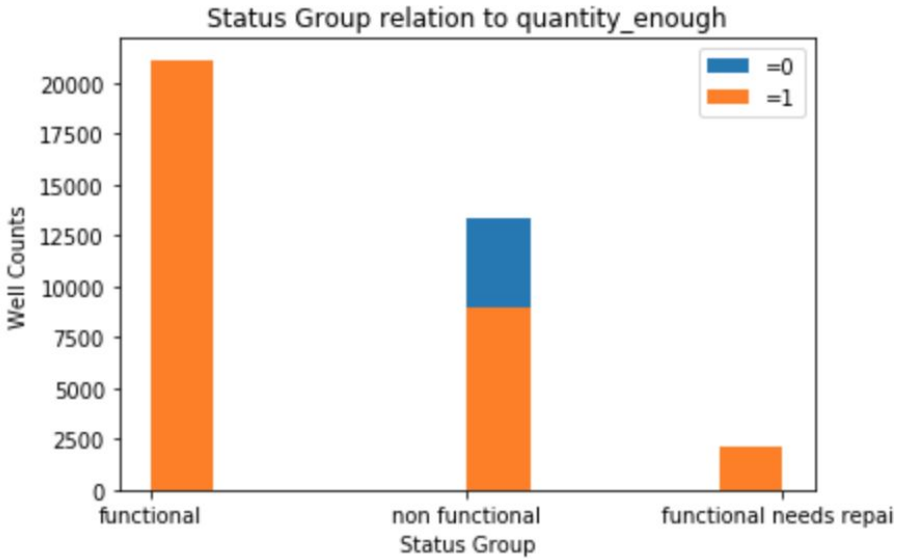
Functional Needs Repair

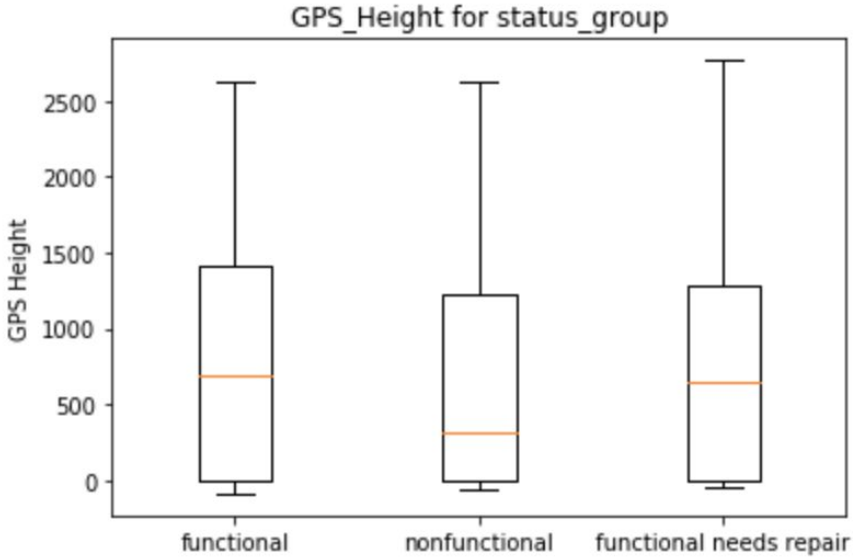


NonFunctional



Quantity_enough feature and GPS Height





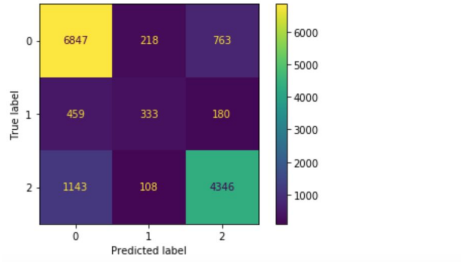
Deciding model to work with

Models compared: Dummy Classifier, XGBoost, Logistic Regression, Random Forest, Decision Tree, and KNN

Metric that was used to decide which model to pick was overall accuracy and recall on Nonfunctional (Category 2).

Models	Overall_Accuracy	recall_2
Dummy_Classifier	0.33	0.33
Logisitic Regression	0.77	0.70
KNN	0.74	0.69
Decision_Tree	0.76	0.76
RandomForest	0.80	0.78
XGBoost	0.75	0.61

	precision	recall	f1-score	support
0	0.81	0.87	0.84	7828
1	0.51	0.34	0.41	972
2	0.82	0.78	0.80	5597
accuracy			0.80	14397
macro avg	0.71	0.66	0.68	14397
weighted avg	0.79	0.80	0.80	14397

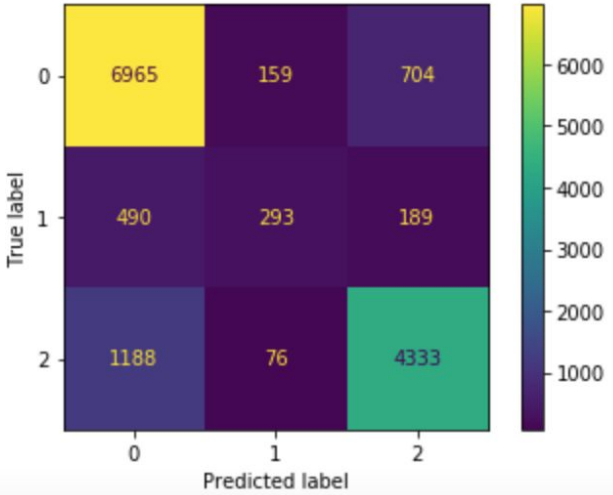


Final Model

Hyper Parameter tuning changed:
- n_estimators: 100 -> 150
- max_depth: None -> 90
- Criterion: ‘gini’ -> ‘entropy’
- min_samples_split: 2 -> 5

Overall Accuracy increased by ~0.005
Recall for Non-functional decreased by 0.01

	precision	recall	f1-score	support
0	0.81	0.89	0.85	7828
1	0.55	0.30	0.39	972
2	0.83	0.77	0.80	5597
accuracy			0.81	14397
macro avg	0.73	0.66	0.68	14397
weighted avg	0.80	0.81	0.80	14397



Summary/Conclusions, Next Steps and Acknowlements

The model that worked best was Random Forest. The final model had a Recall score of 0.77 and an overall accuracy of 0.81. Hyper Parameter tuning did help a little but not by much. The most important aspect of the dataset seems to be location.

Next Steps

Since Hyper Parameter tuning did not affect the model much, I believe the next step is to go back and do some more EDA/data preprocessing.

- Try and do some feature engineering
- Cut down on categorical variables

Acknowledgments:

- Yish, cohort, and tanzanian waterwell group