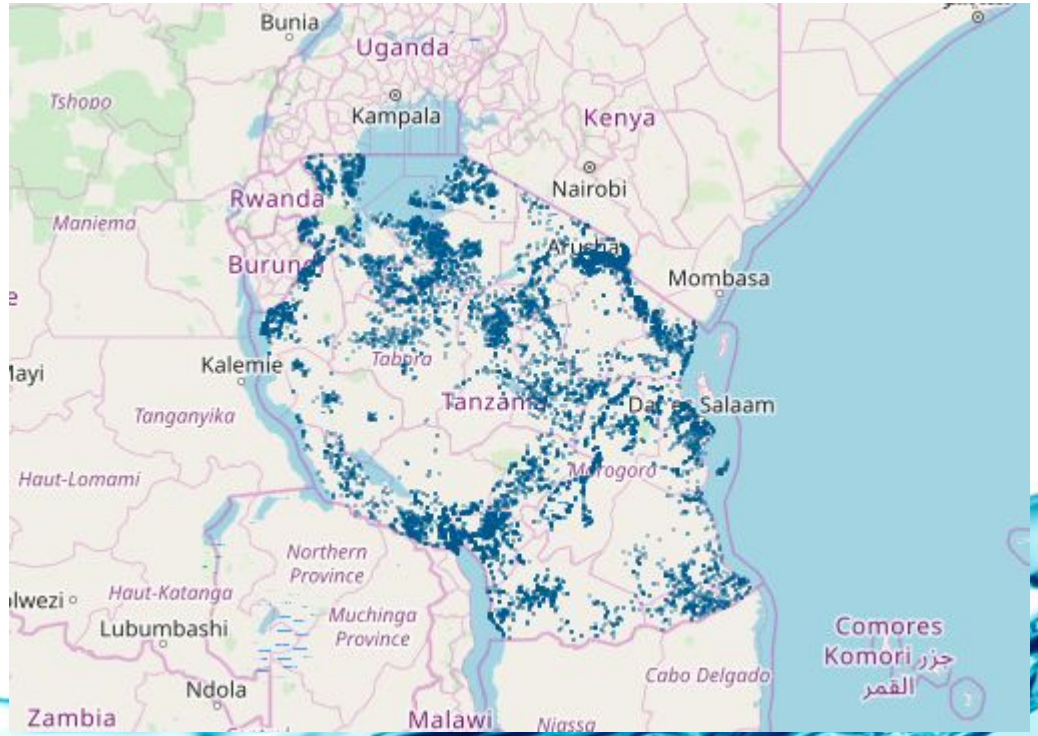
Detecting Water Wells In Need of Repair In Tanzania

By Aaron Abrahamson, Karen Warmbein, and Hunter Ewing

Github

Limited Ability to Monitor/Repair a Fragile Infrastructure

- Only 50% of Tanzanian population has access to safe water
- Out of 59,400 total wells in the dataset, 27,141 in need or repair (46%)
- Our goal is to build a predictive model to indicate the point when a functional well begins to need repairs



Dataset

- “Pump It Up - Data Mining the Water Table” from here
  - 59,400 records with 40 features of water wells in Tanzania
- Large target classification imbalance
  - Ternary classification made binary
    - Wells that needs some repair & broken wells combined
    - What wells need eyes for hands on repair?
- Modeled 17 features (descriptions here)
  - basin, region, scheme management, scheme name, water extraction type, management, payment, water quality, quantity, source, waterpoint type, gps height, longitude, latitude, region code, district code, population, status group
- Dropped 1,812 records during data cleaning

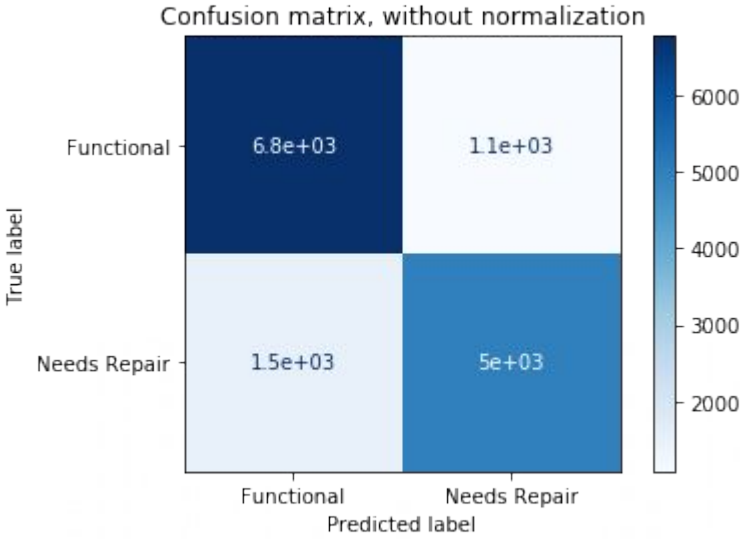
Models

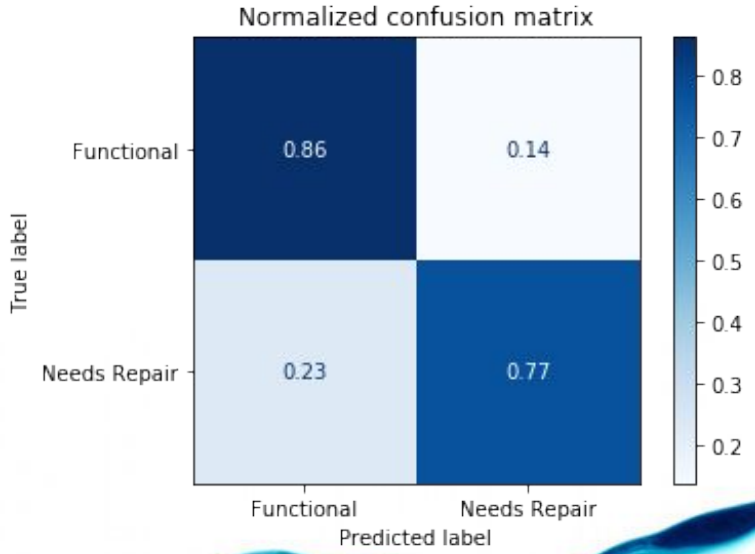
Model	Type	Accuracy	Recall
Dummy	Decision tree classifier Used only numerical features max_depth=5, Gini criterion	40%	----
Model 1	Decision tree classifier max_depth=5, Gini criterion	70.5%	43%
Model 2	**Random forest classifier, 50 trees Hyperparameter tuning, GridSearch Best score: Gini criterion, max_depth=50**	**82%**	**77%**
Model 3	KNN	78.6%	74%
Model 4	Logistic Regression	53.8%	47%
Model 5	AdaBoost	74.8%	61%

Best Model: Random Forest Classifier

- 50 trees
- Hyperparameter tuning
  - GridSearch with:
    - Gini and entropy criterions
    - Limited the depth of the tree to 10, 50 and 100 levels
- Best score: Gini criterion, max_depth=50
- Accuracy: 82%
- ‘Needs Repair’ Recall: 77%

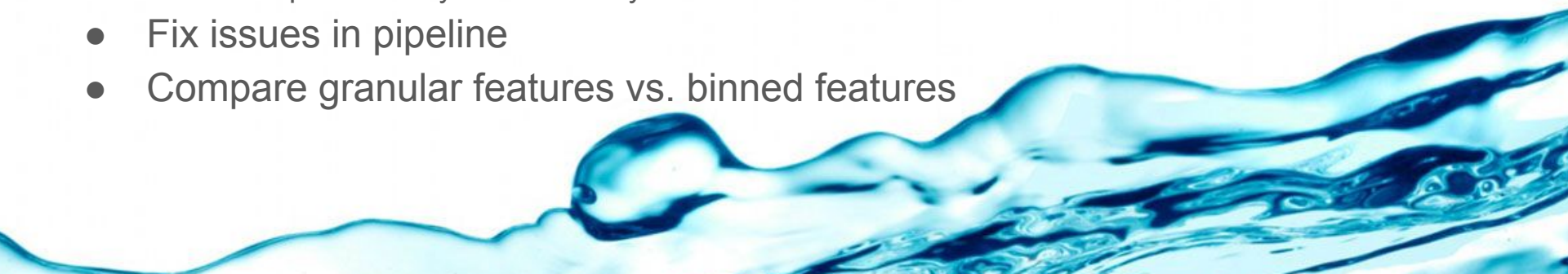
Results


Confusion matrix, without normalization
Functional
Needs Repair
True label
6.8e+03
1.1e+03
1.5e+03
5e+03
Functional
Needs Repair
Predicted label
6000
5000
4000
3000
2000


Normalized confusion matrix
Functional
Needs Repair
True label
0.86
0.14
0.23
0.77
Functional
Needs Repair
Predicted label
0.8
0.7
0.6
0.5
0.4
0.3
0.2

Next Steps

- More feature engineering
  - Replace construction year zeros with the average construction year
    - At least a 2% accuracy increase
  - Replace 0 population with average population; include population in models
    - Assume this is a good predictor
    - More people using the well leads to reduced functionality
- Try more ensemble models
  - Hybrid models
  - Compare ternary versus binary classification models
- Fix issues in pipeline
- Compare granular features vs. binned features



Thank you!

Questions?