Tanzania Water Supply



Background

- 50% of Tanzanians don't have access to clean drinking water
- Groundwater is generally cleaner than surface water which is polluted from sewers and toxic waste
- Sometimes people have to walk miles to closest pump and villages only have one pump

Our Approach

Clean Data

Build Models on Training Data

Predict on Test Data

Data Cleaning

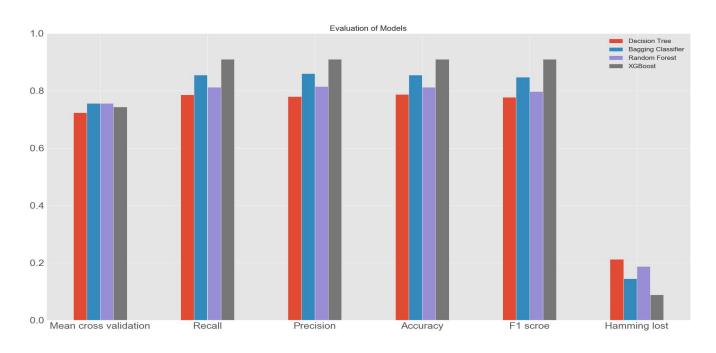
Train data: 59400 records

Test data: 14850 records

- Predict the condition of waterpoint: functional, non functional, and functional needs repair
- Choose features to use
 - Drop meaningless features: id, recorded_by (All by GeoData Consultants Ltd)
 - Use wider range features: larger geographic meaning features, extraction type, The quality of the water
- Adjust unreasonable values:
 - Change NaN to 'others'
 - Altitude of the well is minus
 - The year the waterpoint was constructed is zero
- Avoid overfitting: Only use top 50 values in each features, we set others as 'others'

Models

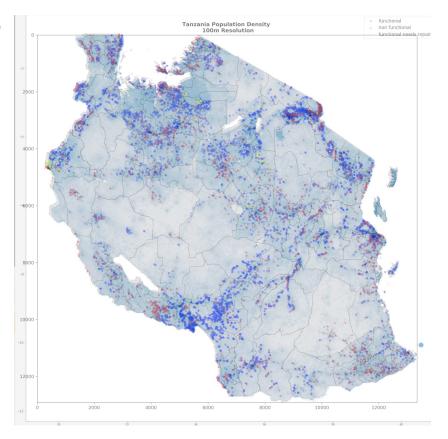
Choose bagging because it had a better cross-validation score and wasn't computationally expensive.



Findings

- The model predicts 64% Functional, 34% Non Functional, 2% Functional Needs Repair in the test data.
- We concluded that population and elevation to be the two main determining factors in pump functionality.
- Most broken pumps in big cities located on the coast with high population density.

Findings



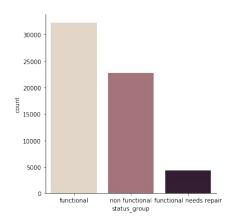
Our recommendation

 Concentrate well repair to cities where it is likely surface water is polluted and pump functionality is imperative for clean drinking water

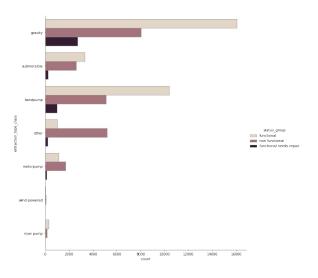


Next Steps:

Address Class imbalance



Better test what types of pumps/wells work better by dropping population and elevation.



Grid Search with XGBoost.

