Analysis and Prediction for Water Well Condition In Tanzania

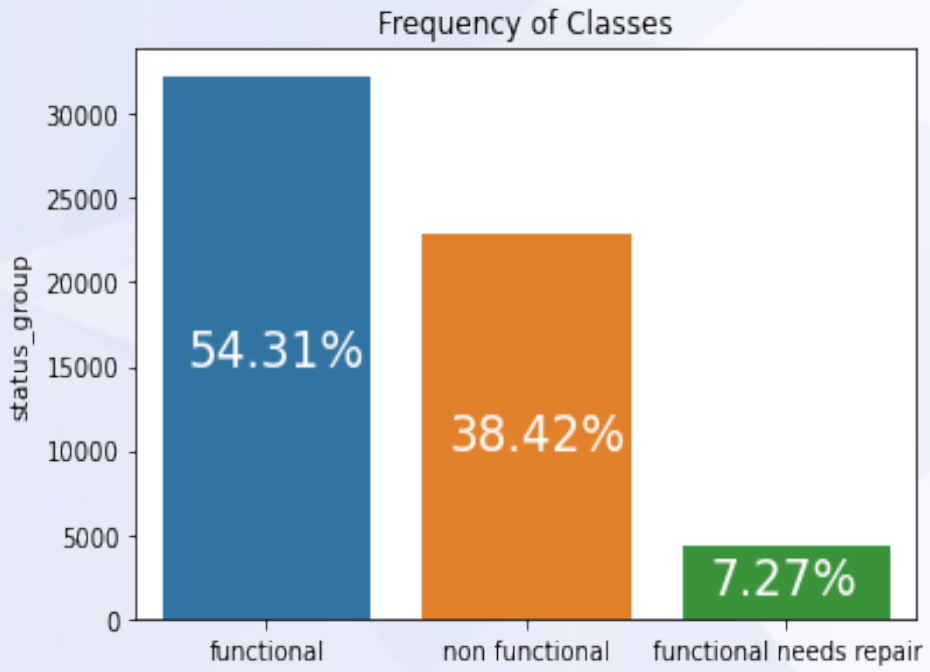
By Alex Billinger

Introduction to Problem

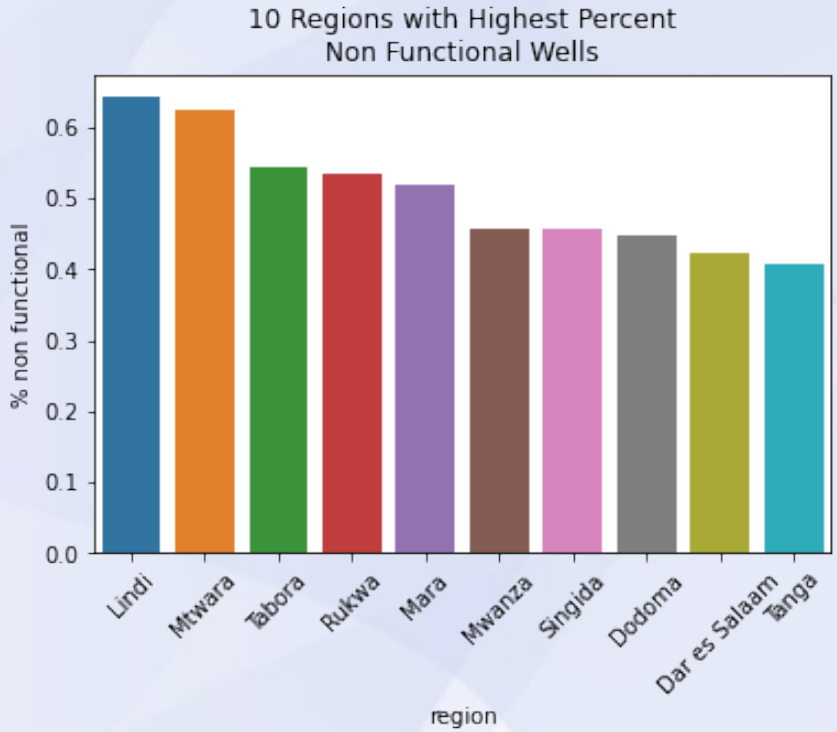
- Track and predict condition of water wells, to provide potable water across country
- Features of data include installing organization, well type, water source, and several categories for location
- Target labels are Functional, Functional needs Repairs, and Non Functional

Are Categories Equal?

- The category of wells that require repairs is very underrepresented in the data set
- Will present additional challenges, as a naive prediction model could be 92% accurate without ever predicting the “requires repairs” category



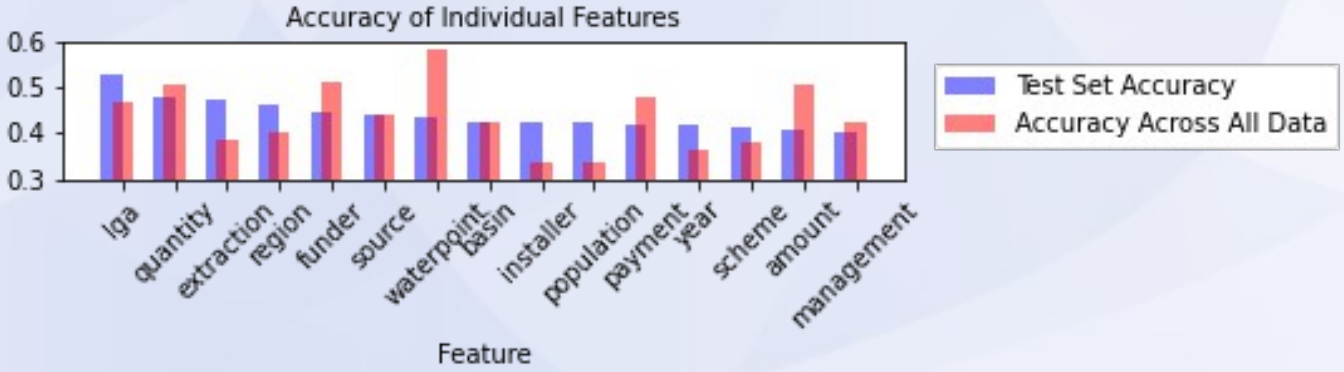
Where to Focus Resources



Where to Focus Resources cont

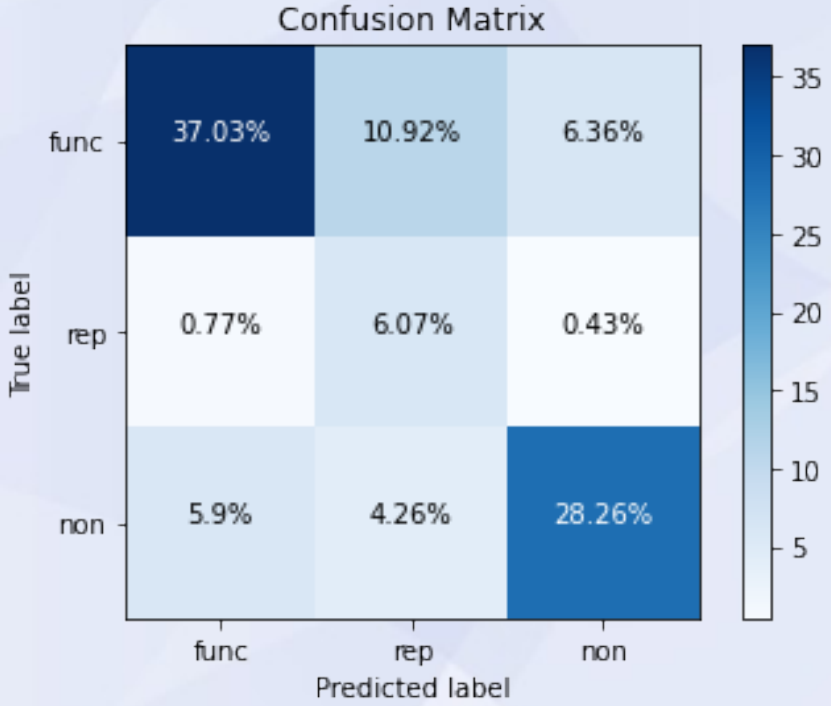
- Compared total number of wells per region to number of non-functional wells per region
- The five with the highest percent non-functional did not start with a huge number of wells, so there are generally relatively few functional wells remaining

Strongest Features



- Removing features with little effect will improve computation time, and allow more optimization
- Accuracy calculated by simple decision tree, used to calculate relative accuracy of different features
- Test set balances target class categories
- Sometimes more accurate across full set by ignoring “needs repairs” category
- No obvious “best” feature, model will require several

Model Performance



- Needs repairs is an important category to track, but hard to make predictions from this data set
- Few false negatives (if repairs needed it will be probably be correctly predicted)
- More false positives than true positives, so many more will be reported as needing repairs than actually do

Future Work

- More data, specifically data over time. Time is quite probably one of the major factors, but we have no data over time. Researching how fast different well types, water quality etc causes degradation would be very useful.
- The “functional needs repairs” category does not have strong predictors, partially because it has so few data points. Requires more data to build a better model

Thank You!