README.md

Tanzanian Well Classification

Author: Micah Shackelford

Business Case

Tanzania has had a problem with available water to the general populace for many years.

The Tanzanian government has hired us to figure out a way to imporve methods in identifying non-functioning water wells.

We will be trying to detect which key features will help up identify the status of these wells.

The Data

Data comes from waterpoints all across Tanzania.

60,000 different waterpoints are included and are classified into “functioning”, “non functioning”, and “needs maintenance”.

Data includes many different features of the wells including geographic location, water source, water quality, pump type, construction year, etc.

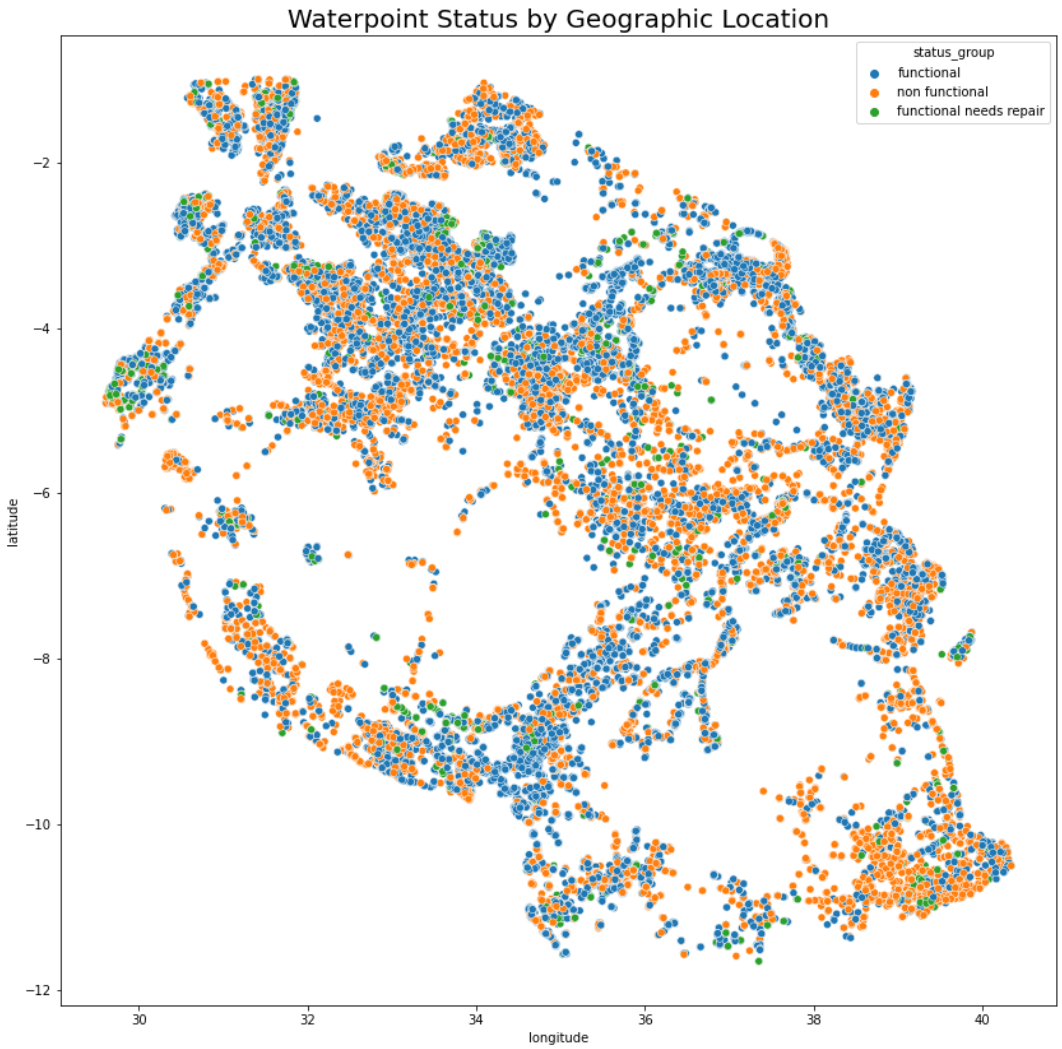
Data Cleaning and Exploration

We filled in missing values, and got rid of duplicated rows.

A lot of our features had over-lapping and sometimes even identical data. We dropped these extra features along with others that were not important for our models.

We binned together some of the data for our categorical features, and we got rid of outliers for our continuous features.

During our exploration we found that location was an indicator of the status of the wells.



We also used scikit learn's One Hot Encoder for our categorical features

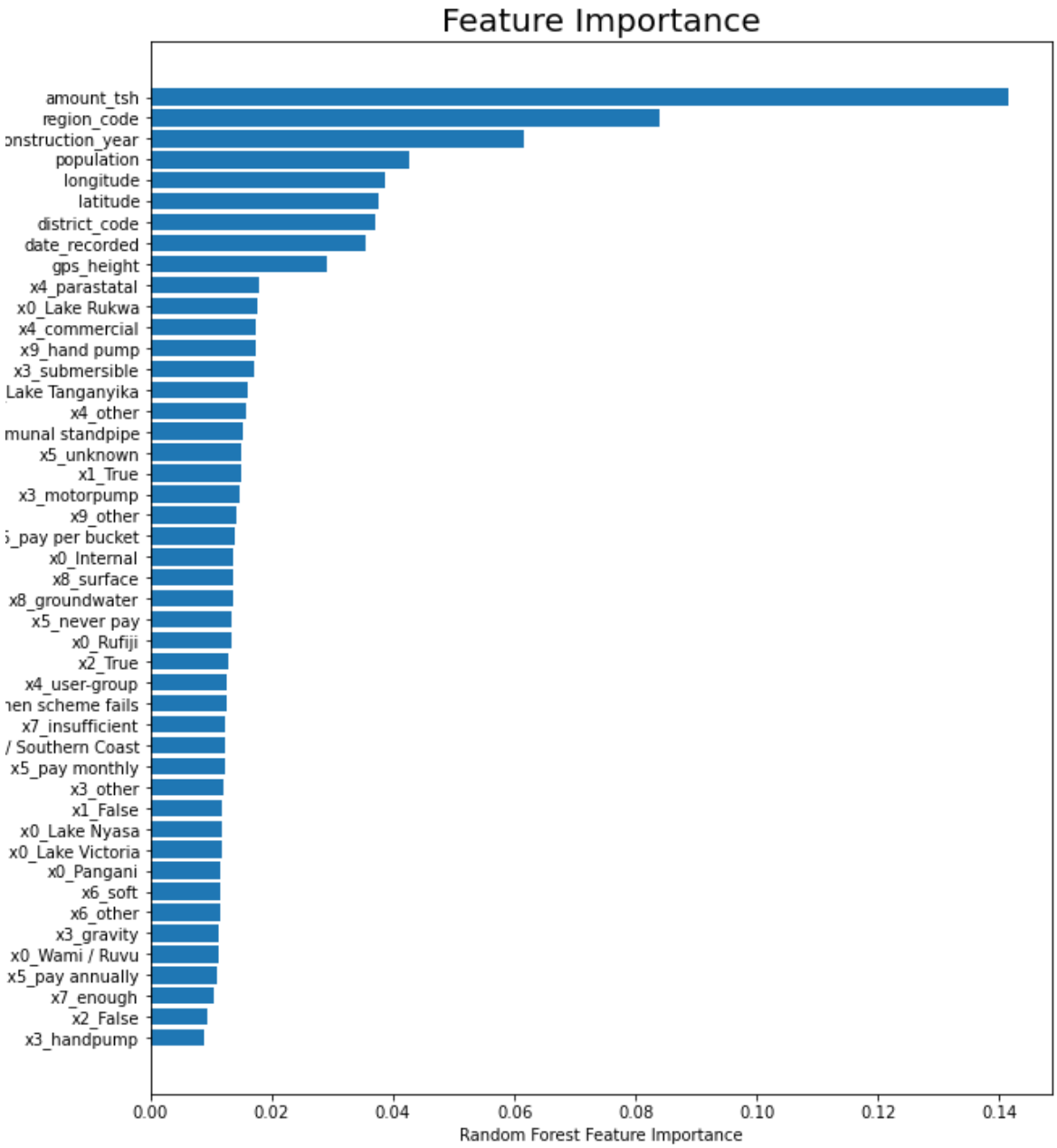
Models

We used three different models for this project. Random Forest, XGBoost, and Decision Tree

XG boost was our best model with an accuracy score of 74.65%

The test data had an accuracy score of 81.52%, so our model did overfit a little bit.

Our model showed that the most important features were, amount_tsh, region_code, construction_year, and population



Recommendations

Look into these water points with low amounts of access to water. Why are these not getting the water they need? If we can give these wells better access to water, we should be able to solve a large portion of the problems.

There is definitely a pattern with location and the status of the wells. We saw this in the visualizations and in our models. Try to find why this is. Is it a problem with the local regulations or something larger?

The population surrounding the wells also seems to be important. A lower population probably means that there are less regulations and maintenance. These wells are still needed though. The government needs to focus on getting these wells back online.

Future Work

We focused solely on the functioning and non functioning wells, since it was the more pressing issue. In the future we need to also better recognize which wells need maintenance, so the gap does not increase.

This data only go up to 2013. Update the data with more recent well information.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%