Tanzania Water Well Challenge

•••

Austin Murray

Overview

This classification problem is a ongoing competition on Drivendata.org

4 million people in Tanzania lack access to an improved source of safe water, and 30 million don't have access to improved sanitation

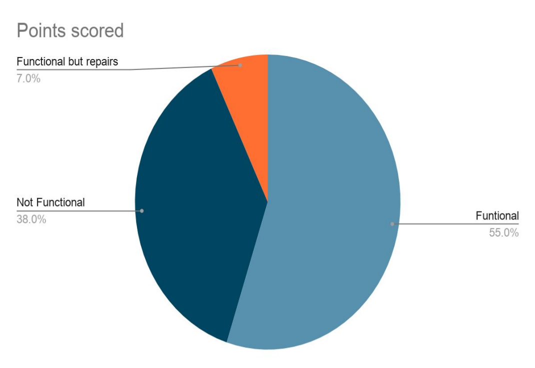
This is an important project as it can help show how data science can help governments with less funding and power to use their resources in a targeted and helpful way to fix problems their populations are living with.

The goal of this competition is accuracy so it is what we will focus on for end goal.

Data

Ternary classification problem

Imbalanced:



- Data was supplied by Taarifa and the Tanzanian Ministry of Water to drivendata.org
- This challenge is to classify functional, functional but needs repair and not functional
- Dataset included 39 independent variables and over 59,000 entries of water wells
- Has not been updated since 2013

EDA

Reduced 39 variables and 59400 rows to 20 variables and 54,612 rows.

Many columns were duplicates or generalized versions of other columns which we took the more specific one since accuracy is our main target.

Filled Nan and 0 rows if the columns were important with “unknown” as to not misclassify data.

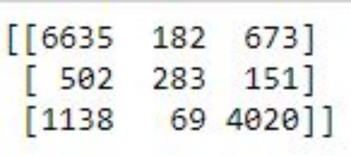
After dumming the categorical we were left with 120 columns to model with.

Models

Final model was random forest that had been tuned with GridSearchCV and Bagged.

Final Accuracy was 80.11%

Confusion Matrix: 13.75% of wells predicted as functional were not.

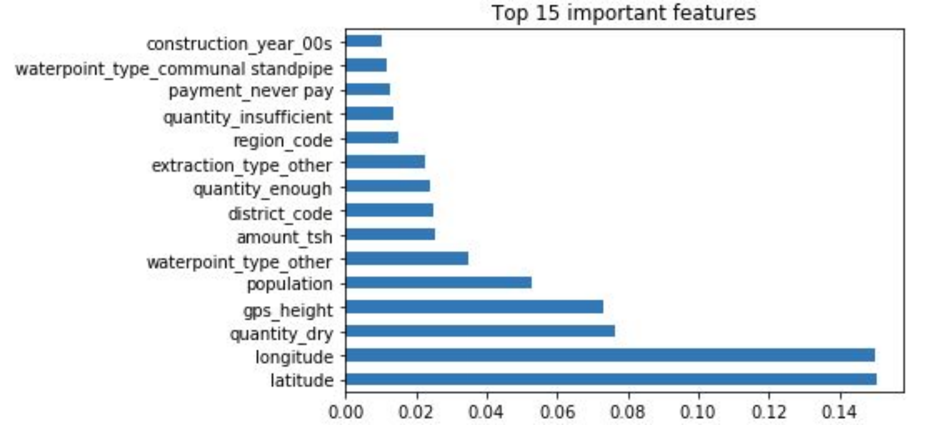


```
[[6635  182  673]
 [ 502  283  151]
 [1138   69 4020]]
```

This shows that was actual vs what we predicted. Misclassified information is a problem still with this model and more eda and tuning will be needed.

Random forest is a collection of uncorrelated decision trees that together are more accurate than one decision tree

Model Cont.

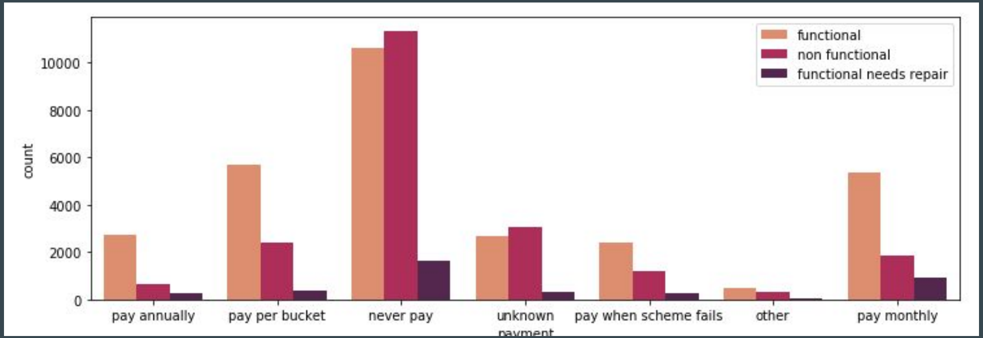


Latitude and longitude being the most important features but having region code and other geographical location low if not on the list shows that these could be adding to the misclassified information.

Conclusion

With a tuned Random Forest we were able to create an accurate model for predicting the ternary classification at 80.11% but with the misclassification at over 13% for just the broken that are labeled functioning the model needs more improvements before government uses this to distribute resources.

Recommendations are to include more information when recording info about the wells. Geographical, weather, and economic situation can play a big part.



Next Steps

- Remove outlier and bin more of the continuous data to see if improvements can be made across the board.
- Latitude and longitude were important features in the decision tree and random forest but when you think about it have nothing to do with a well breaking down so I would like to remove these and gps height and focus more on basin and regions.
- Also want to focus on a bigger break down of the eda to then determine which variables are interacting with each other
- Predict and submit to the challenge

Thank you!

To Flatiron, Taarifa and the Tanzanian Ministry of Water, and you, the wonderful audience here listening.