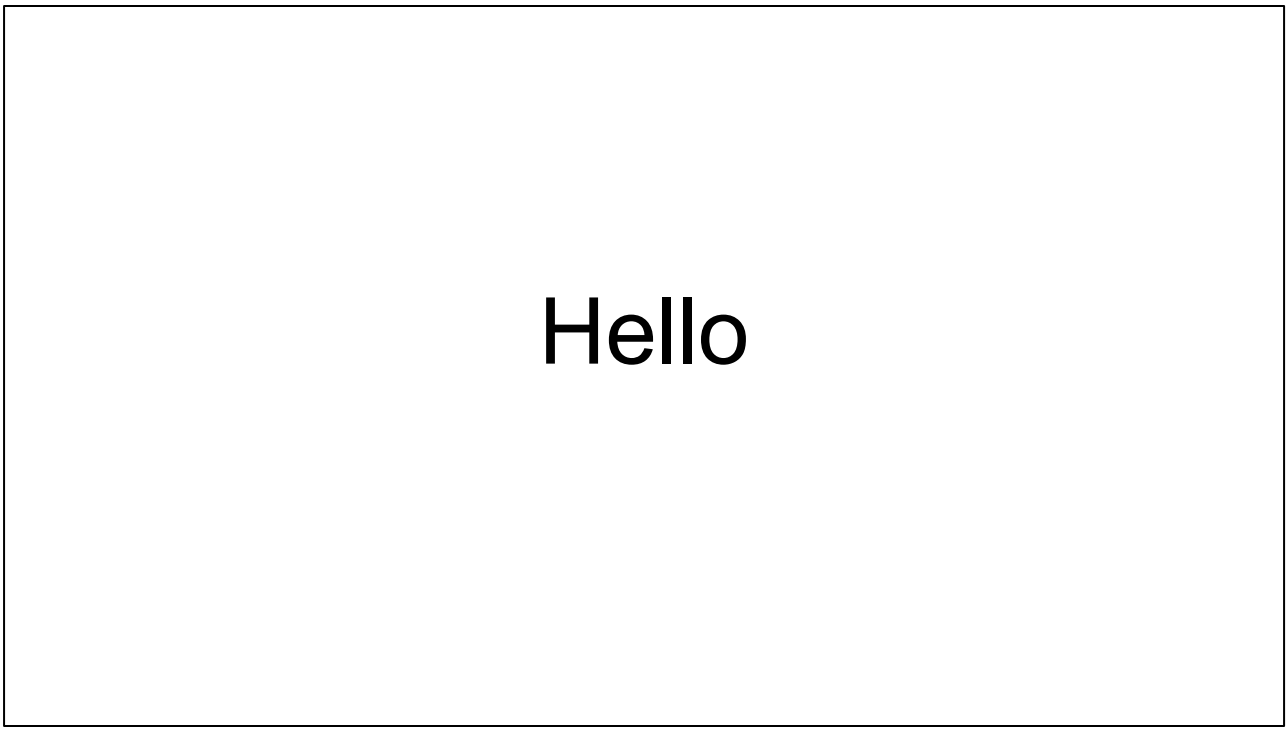


Hello, my name is Mike Erb

Problem Statement

How to More Efficiently Provide Clean Water To The Population of Tanzania

image courtesy of flickr user christophercjensen

I'm here to speak to you about how to more efficiently provide clean water to the population of Tanzania.

Given that the Tanzanian Ministry of Water has limited resources, how can available resources be used to efficiently maintain and expand the water system to provide clean water to the current population of 60 million people?

Specifically...

National Value

Improve Infrastructure Understanding

If we can predict the status of a waterpoint with a high enough accuracy:

1: The Ministry will be able to know the status of any waterpoint without having to make a site visit.

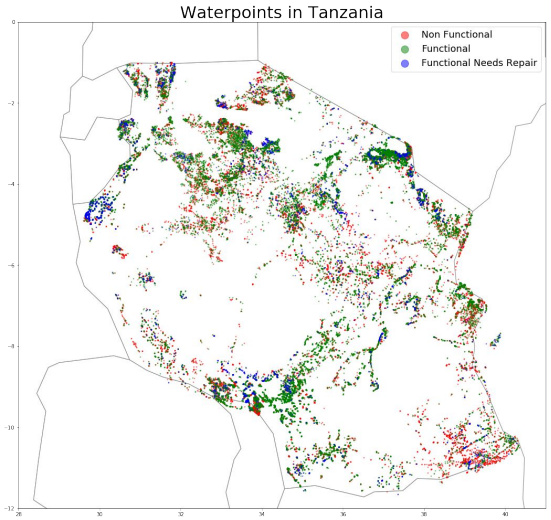
National Value

Resource Reallocation

2: By saving resources eliminating useless site visits, those resources can be allocated elsewhere.

Methodology

Data



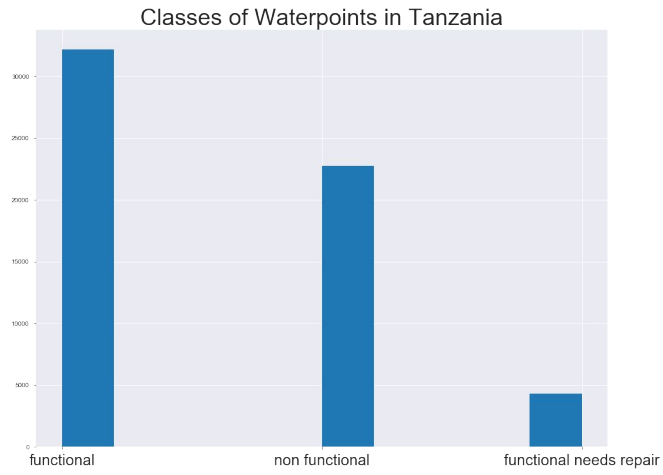
I used the following methodology to create a model to make those predictions:

The data needed to analyze the status of the waterpoints in Tanzania was provided by Taafifa and the Ministry of Water.

It included location, users, water quality and quantity, construction year, type of waterpoint, amongst others of close to 60,000 waterpoints spread throughout the country, which are shown in this map with a color code of its current status of Non-Functional, Functional, or Functional Needs Repairs.

Methodology

Oversampling



The provided data is highly imbalanced, having a majority of the waterpoints functional is good for people, but makes it hard for a model to make predictions.

Using a technique called oversampling, I created new data that is similar to existing data to make a balanced dataset with even numbers of each class of waterpoint.

Methodology

Supervised Learning:

Random Forest Classifier

	Recall			
Model	Functional	Functional Needs Repairs	Non Functional	Overall Accuracy
Logistic Regression - Baseline	0.91	0.01	0.62	0.73
Logistic Regression w/ SMOTE	0.59	0.65	0.61	0.6
Decision Tree	0.74	0.46	0.7	0.71
Random Forest	0.74	0.59	0.73	0.73
XGBoost	0.69	0.57	0.61	0.65

I used supervised learning classification, meaning I knew the class/status a set of waterpoints, and I wanted to train a model that could accurately predict those classes, and more importantly new ones that the model wasn't trained on.

I tested five supervised learning classification models for overall accuracy of their predictions and a balanced recall for each class, meaning I wanted the models to do a good job correctly predicting each class, not just one or two of them.

The type of model that performed the best is called a Random Forest Classifier, which is a collection of decision trees that vote on the class of each waterpoint.

It was able to classify of a waterpoint with 73% overall accuracy, and recall rates of 74% and 73% for the two largest classes (Functional and Non-Functional respectively) and 59% for the smallest class (Functional Needs Repairs)

In this situation, recall can be understood as accuracy within a specific class.

Recommendations

Prioritize Site Visits

Given the success of the model, I have the following recommendations:

Use the model to prioritize site visits.

Priority should be given to maintenance staff sent to waterpoints that are classified to be functional but in need of repair or non functional.

Recommendations

Expand Infrastructure

Use resources that would previously have been budgeted for site visits to expand the water infrastructure.

The model can be used to classify the nearby waterpoints to prioritize underserved areas.

Recommendations

Solicit International Aid

Use the classification model as a way to show international aid organizations that their investments will be worthwhile because the Ministry of Water will be better able to maintain current and future infrastructure.

Future Work

Investigate Problematic Misclassifications

		Functional	Functional Needs Repairs	Non Functional
	Functional			
True	Functional Needs Repairs			
	Non Functional			
		Predicted		

BAD	OK	GOOD

We were able to get an accuracy of 73%, but can we do better and how? This question brings me to future work.

In our situation we have three classes (functional, functional needs repairs & non functional) for our waterpoints that we want to correctly classify.

In this table green indicates correct, yellow indicates wrong, but ok, and red indicates wrong, but not ok. We need to focus on correcting the misclassifications in the red zones.

For example if a waterpoint is non-functional but is predicted as functional needs repairs, it's a yellow square. That is not much of a problem because the maintenance staff would visit the site and just find they had to do more repairs than anticipated.

Compare that to a Non-Functional site being misclassified as functional, a red square. In this case the waterpoint would not be visited and would remain non-functional.

Future Work

Maintenance Records

While the current models does a good job, going forward we will need additional data to keep the model up to date and improve its classification power.

One way to do this will be gathering historical maintenance records and implementing a structure for collecting them in the future so that they can be incorporated into the model.

Future Work

Useless Waterpoints

The Random Forest Classifier found that the most important feature for classifying a waterpoint was whether it was dry, meaning regardless of the functionality of the waterpoint there was no water.

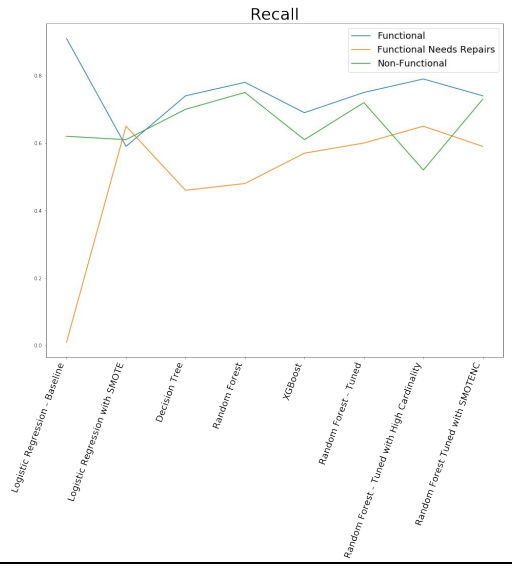
In the dataset, dry waterpoints were almost exclusively non-functional.

An attempt should be made to train the model to further classify non-functional waterpoints as to visit or ignore. Dry is an obvious characteristic, but a classification model may be able to determine if there are others or combinations of others, to assist in further optimizing waterpoint maintenance operations.

Thank You

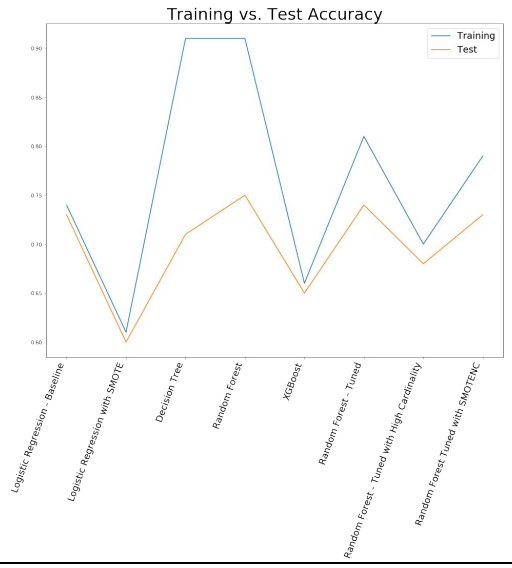
Appendix

Recall for each class for each classification model and three tunings of the Random Forest Classifier.



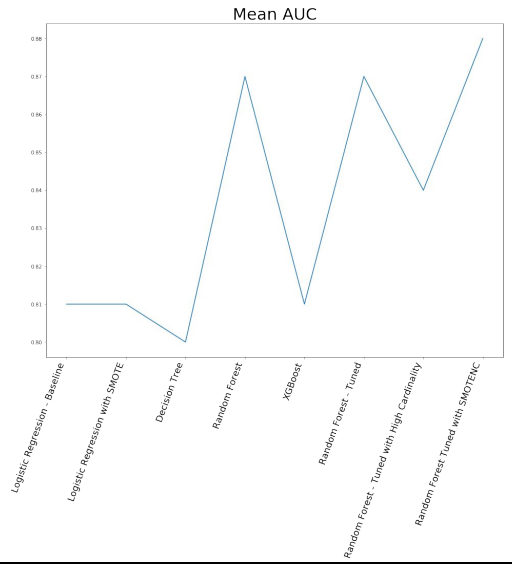
Appendix

Training Vs. Test Accuracy for each class for each classification model and three tunings of the Random Forest Classifier.



Appendix

Mean AUC, a secondary performance metric in this case, but one that reflects, the primary metrics of accuracy and balanced recall.



Appendix

Feature Importances as determined by the best performing Random Forest Classifier