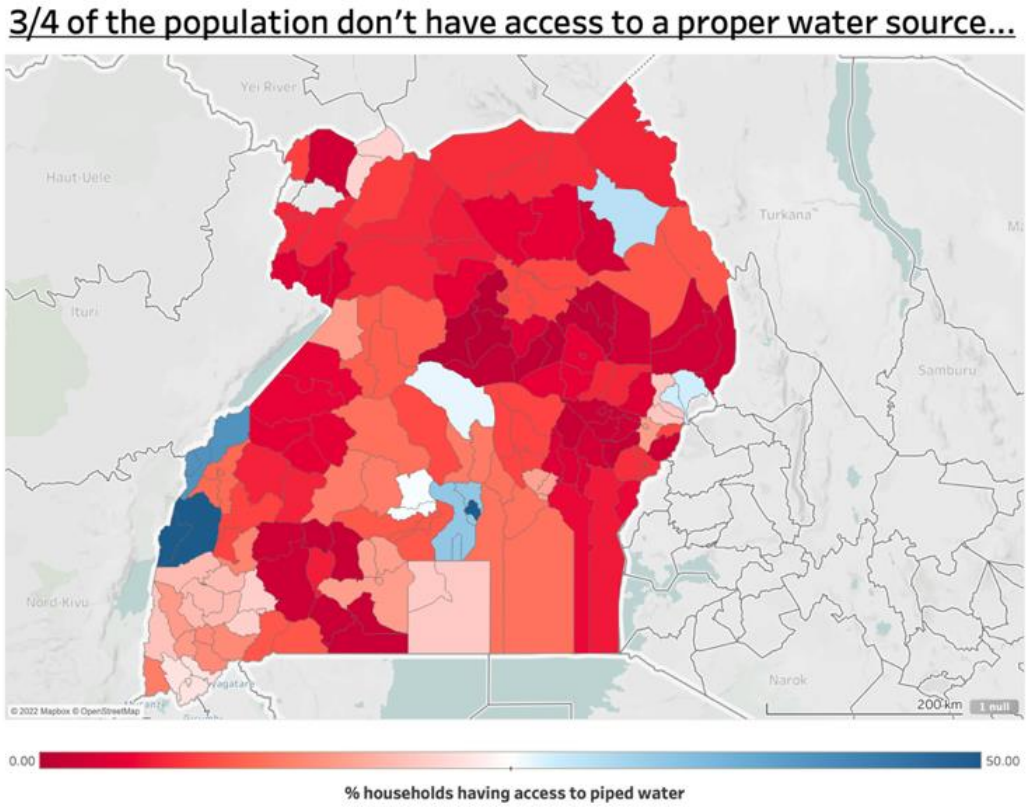
Water Infrastructure in Uganda
Data Science – Thomas Adler – August 2022

1. Why is access to water so important?

Water is one of human's most basic needs. It helps us keep healthy, functioning and producing other basic needs such as food and shelter. As a result, access to water heavily impacts the development and prosperity of a country. Health, education, income, production, and inequality are all dependent on access to water. This is an especially important problem in Uganda, where ¾ of the population do not have access to piped water and half of healthcare facilities do not have access to running water. This poverty and inequality also leads to political instability and violence, more than 31,000 people have been killed in conflicts in Uganda alone since 1997. I want to help solve the issue of water access. Uganda has many water points across the country, and they should be enough to service every single citizen. However, too many of them break down and are not repaired. I want to build a model that will better predict which water points have, or are at risk, of breaking down, so the Ugandan government and local public institutions can better monitor, improve and repair crucial water points.



2. How can I solve this problem using Machine Learning?

This is a classification problem. We need to predict whether a water point is working or not, at any given time. There has been one attempt to do this with a limited dataset, with only very specific information on the water point itself. I am adding value because I will be introducing a wider range of economic, social, public features as well as information on local conflict. I expect my technique to work better because I will be better capturing what makes a water point functioning or not. My problem statement is:

"How might we use Machine Learning to predict whether a water point has broken down in order to repair it quicker?"

3. Which datasets did I use?

I used three datasets. The first was data on every single water point in Uganda. Thanks to the Ugandan's government great statistical capacity, this is up to date and contains a lot of information on the technology, status and characteristic of the point. I used the API from the Water Point Dataset Exchange (WPDx), which is an NGO that collects that information from governments. The second was

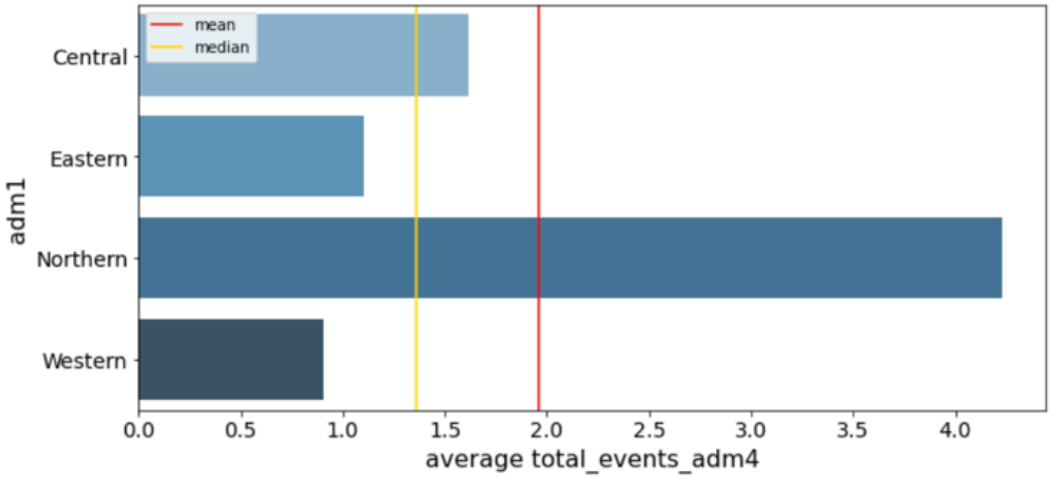
regional demographic data from the Bureau of Statistics (BoS) of Uganda. There is information for every regional level in the country and I had to manually download those. Finally, I used the API of another NGO, Armed Conflict Location Event Data Project (ACLED) to get every single conflict/event that happened in Uganda, since 1997. Uganda has been involved in wars with external countries (notably the Congo) as well as having a lot of armed rebel groups in the country, creating instability and violence. Note that Uganda is not a very democratic country.

4. How did I prepare the data?

For the conflict data, I created summary variables by region (total number of fatalities/events). For the demographic data, there was not much to be done, apart from converting all numbers to percentages. For the water points, I had to binarize and convert to numeric values a lot of information as well as drop columns which had too many missing values. Then I joined those three datasets on sub-counties. I ended up with each observation being a single water point, with specific information on its characteristic as well as regional demographic and conflict information. Finally, I did some initial feature engineering and dropped variables based on certain violated assumptions (multicollinearity, linearity, variance…).

5. What did EDA tell us?

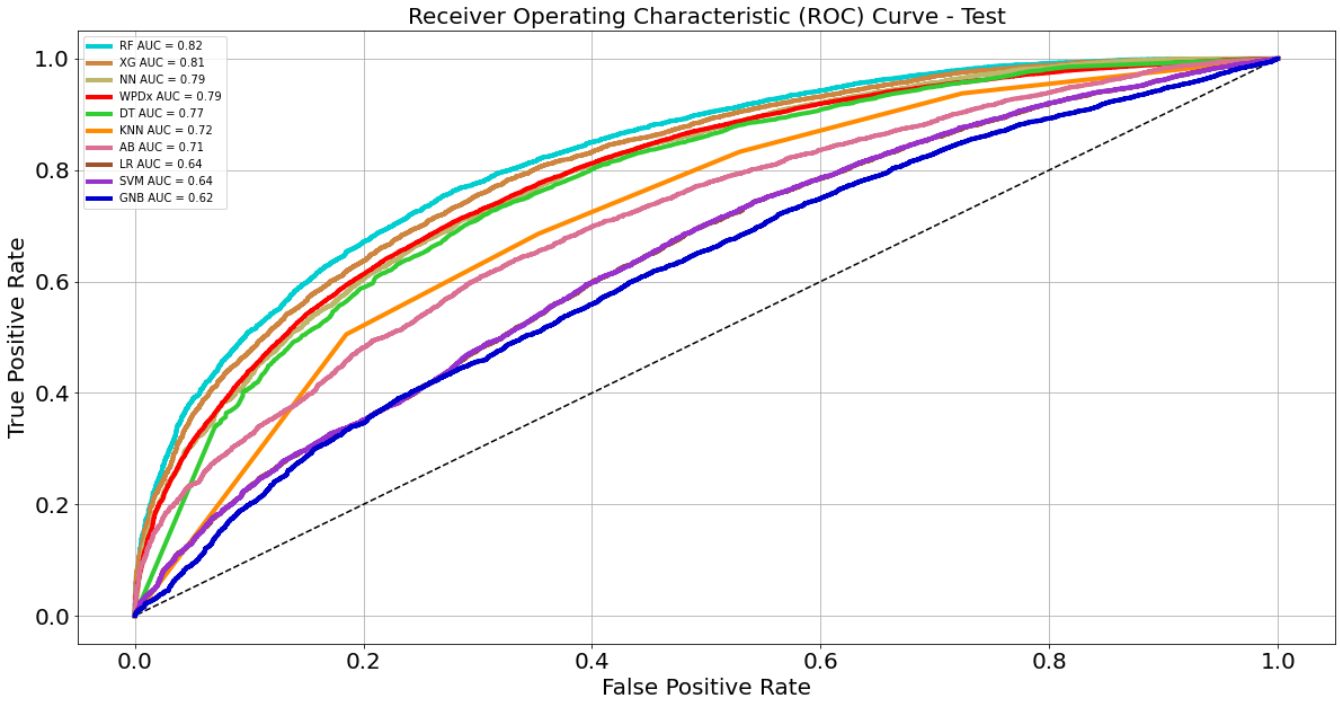
The most important information from the EDA process was the regional differences in the country. Broadly speaking the Central region (including the capital, Kampala) is the urban centre and most developed part of the country. The Northern region is the least developed and the region which has suffered the most from violence.



In addition, I identified that 2/3 of points are not of complex technology, 2/3 have been installed after 2006 and 3/4 are managed by public bodies. Overall, 1/5 of all water points are not functioning at any given time.

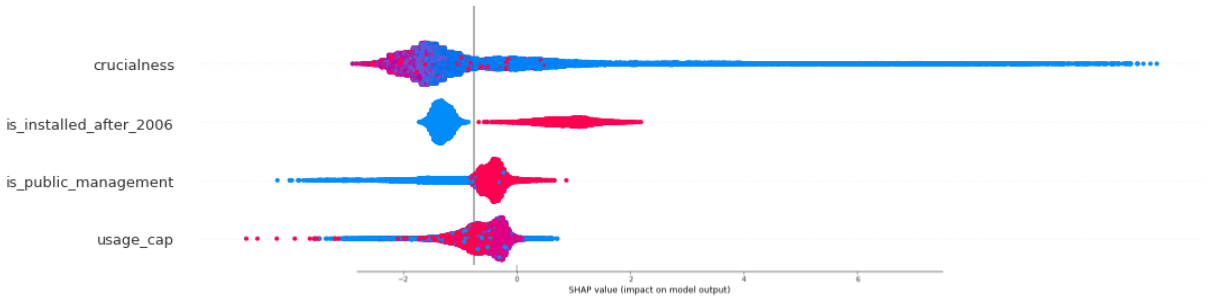
6. What was the modelling process?

I decided to run as many ML models which could lend itself to a classification problem as possible. The process was similar for each one. I first resample the data to increase the share of my minority class (non-functioning points) and scale the data when required. Then I run a baseline model, with all the default parameters. I evaluate its confusion matrix and various accuracy metrics. Then I visualise the impact of the different hyperparameters on the accuracy score, to get a rough idea of where to focus my grid search. After that, I run a grid search or random (depending on how expensive the model is) cross validation. I also include the option of dimensionality reduction using PCA. Having found the optimal model, I rerun my model with the tuned hyperparameters and compare the confusion matrix, accuracy metrics and ROC curve with my baseline. I choose the best model based on their recall score for non-functioning points on the test set. With that final model, I visualise the feature importance, export the metrics and dump the model. In the end, my best model was XGBoost as it had the highest recall score for non-functioning points, as well as good other accuracy metrics, a shorter prediction time than most models, the second best AUC and a relatively small overfit on the training set.



7. <u>What are some key findings and subsequent recommendations?</u>

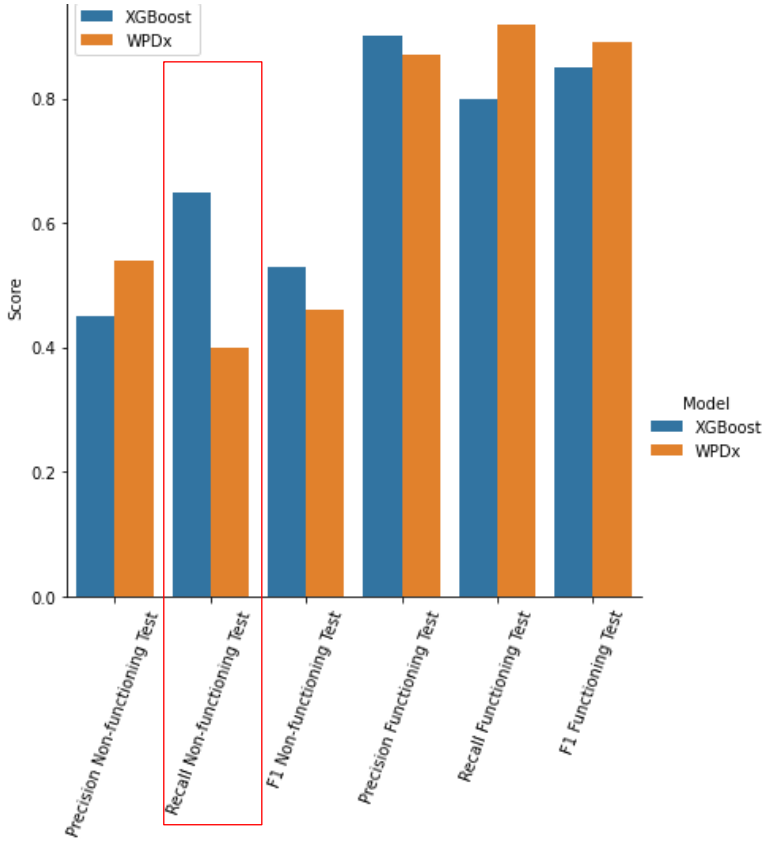
Using Shapley values for our XGBoost model as well as the feature importance of other models, I find that the best predictor for predicting a water point functioning is how crucial that water point is. Crucialness is the proportion of the local population that depends on that water point. In addition, the usage capacity of a water point, whether it was installed after 2006 and whether it is publicly managed are all strong predictors of water point functionality. This matches some of our initial hypotheses. One surprising finding is that local conflicts did not have a strong predictive effect on our outcome. This is probably because the initial reason why there is conflict is poverty and under development, something I have attempted to capture in other variables.



From these findings, I would recommend that public bodies focus on those points which are the most important to local communities. Intuitively, this is where the most impact can be made as repair and improvement work will impact the most individuals. However, a word of caution to keep servicing more rural and less crucial water points. These smaller communities should not be left behind and lack of water access will still heavily impact them. Not giving these communities enough attention will create inequalities and put the overall development and stability of the country at risk.

8. What is the impact of our model?

Our model greatly outperforms the current, WPDx, model. It identifies 50% more non-functioning points, resulting in $USD 28m in economic benefits generated and 13,000 deaths prevented each year from reliable access to clean water.



9. What are potential next steps?

An obvious way to improve our model would be to include more data. This can be hard to find, but data on health outcomes, especially the mortality rate due to water-inducing sickness would be extremely helpful to understand if a water source has been damaged or contaminated (akin to the famous Jon Snow experiment in London). We could also include data for other countries, or even test our model in other water-deprived countries and test the external validity of our model.

Also, I have done a relatively extensive step of feature engineering before running the model, in an attempt to introduce more qualitative analysis. It might make sense to run our models with all the available features, and introducing more aggressive regularization, to see if the models choose to drop the same variables as I did and if it gets better accuracy metrics. In addition, it would be worth spending more time on neural networks and test a wider range of complex architectures to look for improvements.

Finally, it should be said that my classification problem and the features included might not lend itself to more complex models. It might be that the relationship between my features and my outcome is relatively straightforward and does not need more complicated steps. Anyhow, the model that best enables the citizens of Uganda to meet their basic needs should be prioritised.