Unlocking the Flow : An Analysis of Tanzanian Water Wells

BY WILLIAM OMBALLA

INTRODUCTION

Amani a resident of a village in Tanzania fully relies on a communal water well for access to water. However, she is not sure if she will be able to get water today. The well has been malfunctioning and it's a matter of time before it stops working.

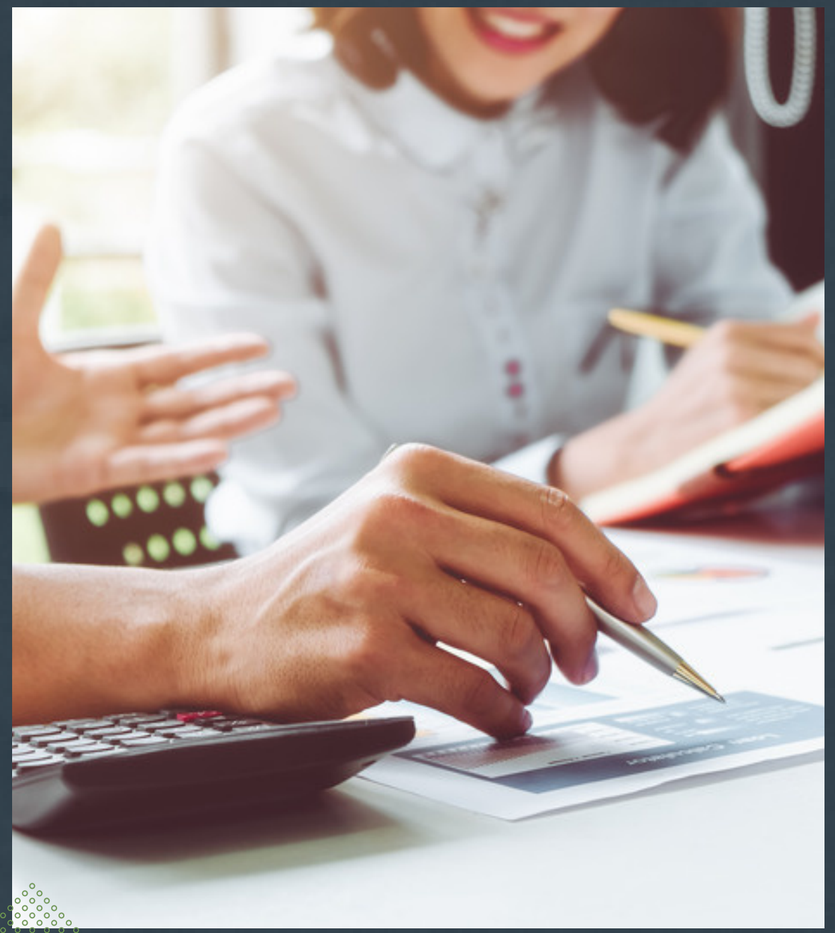
Problem Statement:

The majority of Tanzanians rely on water wells for access to water. However, the reality is that half of the wells are either in need of repair or are not functional at all.

Solution:

By analyzing the status of the current wells, insight can be drawn to help infer why some of the wells are not functioning and possibly create a model that can predict the status of a well based on the collected data enabling better management of wells across Tanzania.

BUSINESS & DATA UNDERSTANDING



Stakeholders

The stakeholders involved in water well interventions in Tanzania include government agencies, non-profit organizations, private companies, and local communities.

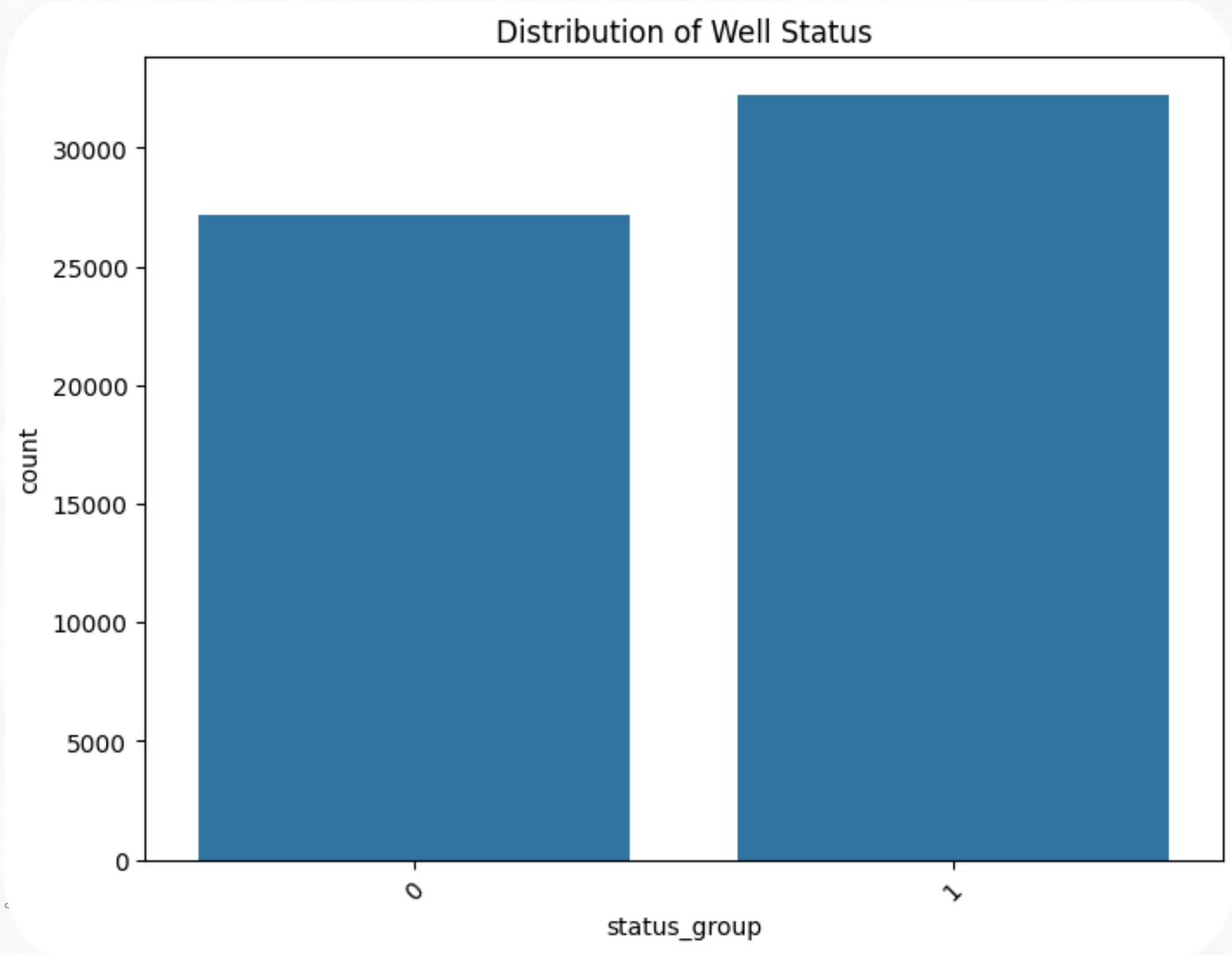
Data Source

The data source used was curated by an organization named Taarifa in Tanzania.
The data features a target class that is ternary i.e. has 3 classes.

1. Functional
2. Functional needs repair
3. Not functional

For the scope of this project, the Functional Needs Repair class has been merged with the ‘Not Functional’ class. This is to enable binary classification.

DATA ANALYSIS



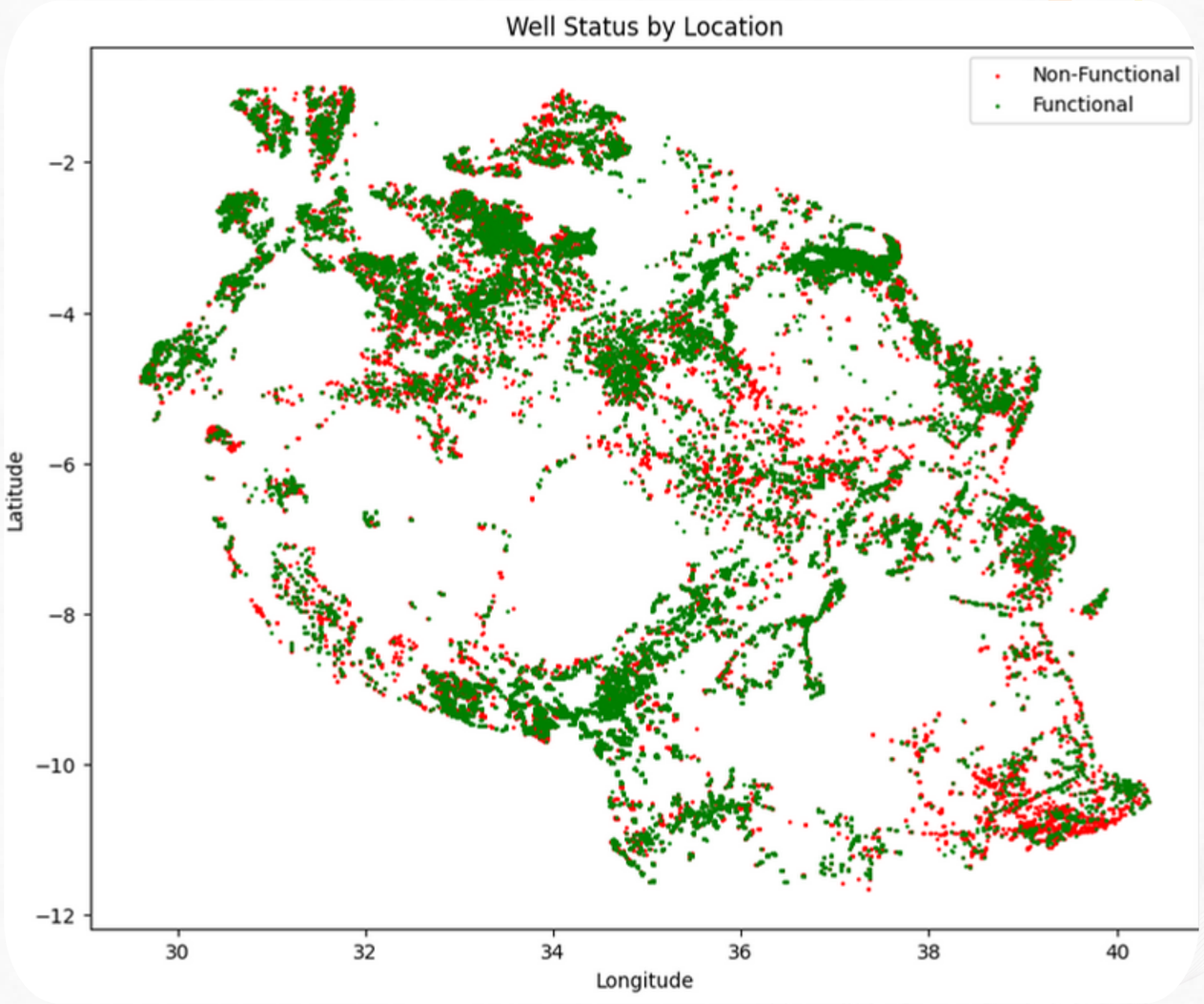
Class Distribution

The first analysis would be to confirm the class distribution of the target class since we merged two classes. The distribution shows that there is no class inbalance generated by the merge.

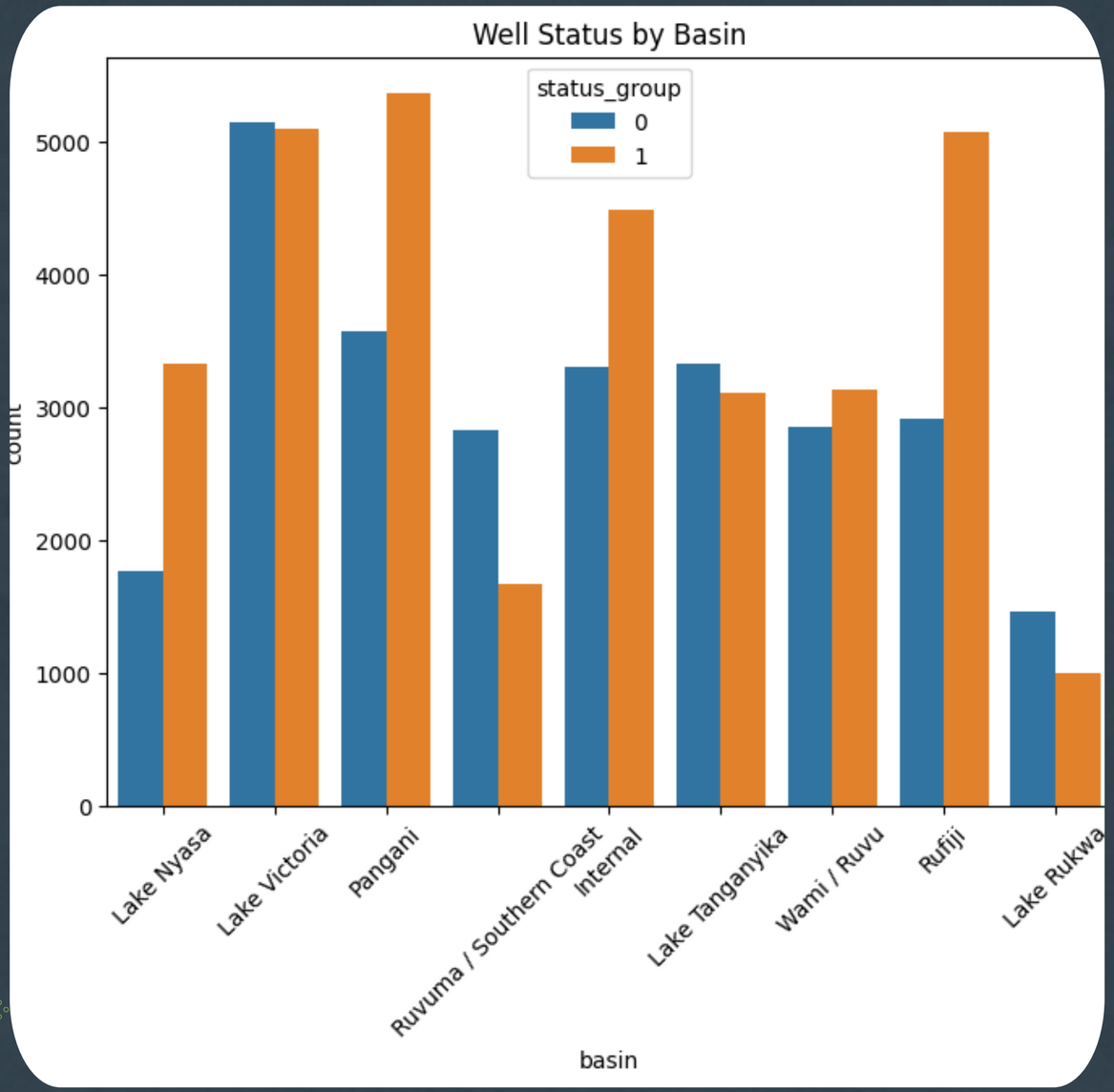
DATA ANALYSIS

Well Status Distribution

The graph helps show the distribution of the target class based on the longitude and latitude. As can be seen, there are no areas concentrated with one class.



DATA ANALYSIS



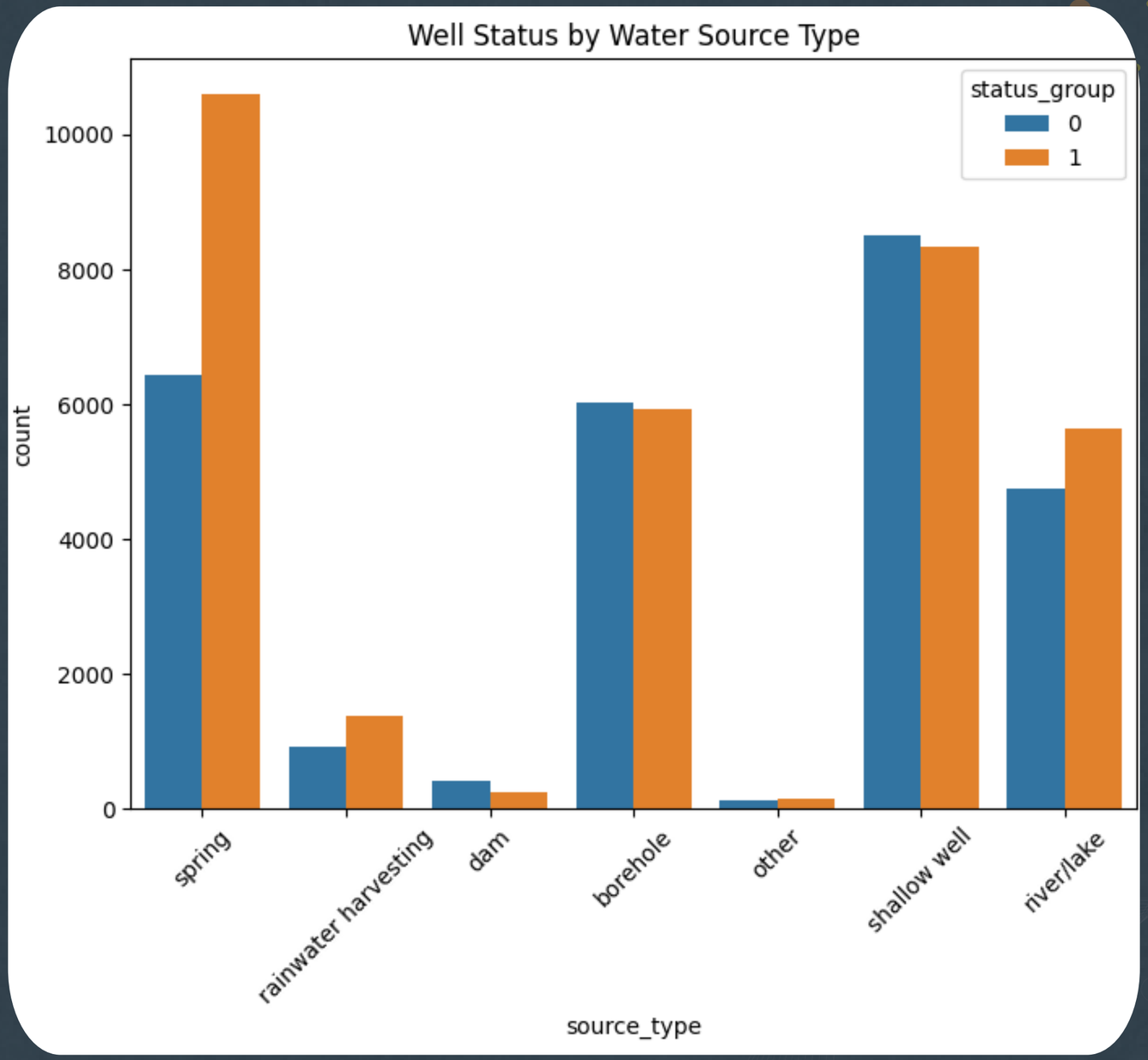
Well Basin

Grouping the target class by the water basin we can see that Lake Nyasa, Rufiji, and Internal all have more functional than nonfunctional wells with Ruvuma, Lake Tanganyika, and Lake Rukwa having more nonfunctional lakes than functional

DATA ANALYSIS

Well Water Type

Spring water type seems to be the highest with functional wells. We can also see that boreholes and shallow wells can be improved upon.



MODELLING

For the modeling, I went with three models to gauge the performance of each on the dataset. Both the Logistic Regression Model as well as the Random Forest work well with binary data. (Data was converted to binary).
Though the Logistic Regression was okay, achieving an accuracy of 73.5%, the Random Forest was able to surpass that and achieve an accuracy of 81.5%. This prompted me to run the Random Forest again this time with tuned hyperparameters to squeeze out the performance of the model. This resulted in a slight increase in accuracy to 82.0%.

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Logistic Regression	0.735	0 - 0.77 1 - 0.72	0 - 0.59 1 - 0.86	0 - 0.67 1 - 0.78
Random Forest	0.815	0 - 0.81 1 - 0.82	0 - 0.77 1 - 0.85	0 - 0.79 1 - 0.83
Tuned Random Forest	0.8203	0 - 0.84 1 - 0.81	0 - 0.75 1 - 0.88	0 - 0.79 1 - 0.84

EVALUATION

Although the model accuracy is good, there is a general trend that the recall is relatively low compared to other metrics.

This can be attributed to the way we converted our model from ternary to binary. We combined all “Functional wells needs repair” classes to “Non-Functional”. This can suggest that some of the features of these classes are different hence the model cannot learn well.

RECOMMENDATIONS.

Increasing the number of boreholes and shallow wells or fixing the currently available ones can help reduce the Non-functional wells and possibly increase the water supply.

Ruvuma, Lake Tanganyika, and Lake Rukwa have a high number of nonfunctional wells that should be looked into and repaired.

More accurate data on the installer of the wells should be provided as this can help identify contractors with durable wells that can last longer before breakdown.

NEXT STEPS.

Further research can be done on the dataset to find out the wells that are around the most population. This can help identify areas that need urgent attention.

Modeling a ternary classification model can also assist in correctly identifying the status of the well and will likely remove some ambiguity generated by combining the two classes.

Collaboration with relevant authorities to help predict well malfunctions as well as repair the ones in critical condition can help prevent water shortages.

THANK YOU

FOR YOUR ATTENTION