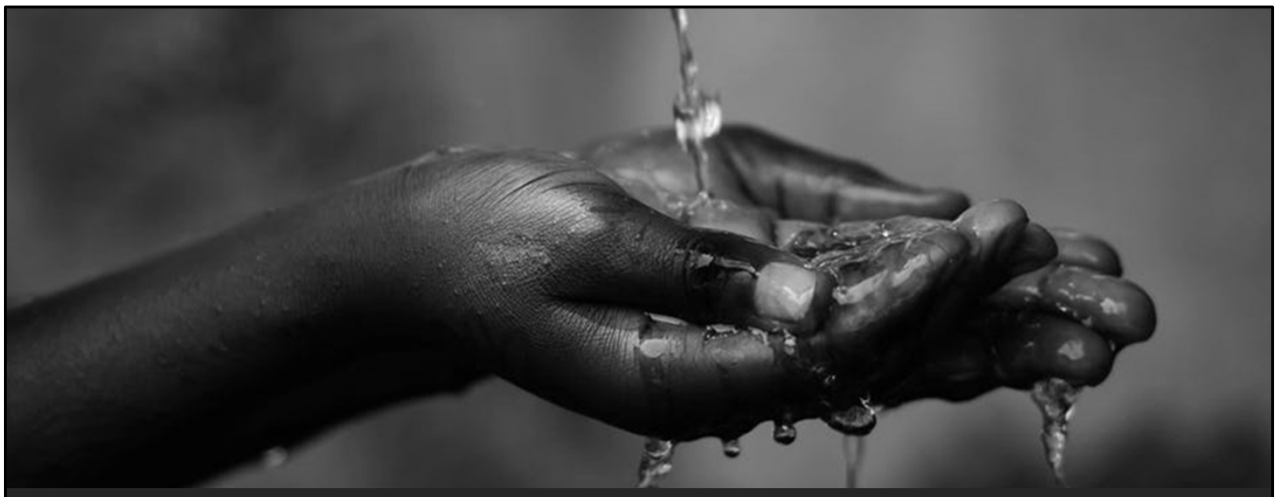

Water Analytics

Clean and Sustainable Water Supply in Nigeria

Credit: Photo from https://theconversation.com/mozambique-water-project-insights-into-supply-and-use-in-a-peri-urban-area-115171

Introduction

- Clean and accessible water critical to human health
- Healthy environment, sustainable economy, reduces poverty and ensures peace and security
- Over 40% of global population do not have access to sufficient clean water
- By 2025, 1.8B people will be living in areas with water scarcity
- Up to 80% of illnesses in developing world are linked to inadequate water and sanitation

Ensuring the availability and sustainable management of water and sanitation is one of the Sustainable Development Goal identified by the United Nations. However, providing clean potable water to rural communities in many developing countries remain challenging due to water scarcity and poor water quality. The lack of clean water poses a threat to a community's health, safety, and economy.

Objective

- Address the issue of providing clean and water supply to the rural community
- Build model to predict Non-Functional water points in Nigeria
- Model built using logistic regression and decision tree
- Model should be relatively simple, with good performance and relevant variables

The objective of the analysis is to build models to predict the reliability of water points in Nigeria, so that preventive maintenance can be carried out on urgent Non-Functional systems timely. The models will also be compared to identify a model with a suitable accuracy and not overly complex.

Variables Considered

- Local population with 1km radius
- Served population
- Water source
- Water technology
- Administrative area
- Payment system
- Subjective Quality
- Urbanisation
- Latitude
- Longitude
- Distance to primary road
- Distance to secondary road
- Distance to tertiary road
- Distance to city
- Crucialness

The dataset is obtained from the Water Point Data Exchange (WPdx) Data Repository. The repository collects water point related data from rural areas at the water point or small water scheme level, based on WPdx Data Standard, and shared on a cloud-based data library. https://data.waterpointdata.org/dataset/Water-Point-Data-Exchange-Plus-WPdx-/eqje-vguj/data. After data cleaning, these are the variables extracted to be used for the model

Models

		Training		Validation		Testing	
		Rates		Rates		Rates	
Method	Actual	Non-functional	Functional	Non-functional	Functional	Non-functional	Functional
Predictor Fit Nominal Logistic	Non-functional	0.705	0.295	0.698	0.302	0.703	0.297
	Functional	0.343	0.657	0.339	0.661	0.335	0.665
Stepwise (Forward)	Non-functional	0.699	0.301	0.694	0.306	0.701	0.299
	Functional	0.343	0.657	0.342	0.658	0.335	0.665
Stepwise (Backward)	Non-functional	0.702	0.298	0.699	0.301	0.705	0.295
	Functional	0.344	0.656	0.341	0.659	0.334	0.666
Predictor Partition	Non-functional	0.754	0.246	0.742	0.258	0.741	0.259
	Functional	0.338	0.662	0.341	0.659	0.336	0.664
Predictor Bootstrap Forest	Non-functional	0.790	0.210	0.757	0.243	0.761	0.239
	Functional	0.308	0.692	0.327	0.673	0.321	0.679

- 5 models generated
  - 3 logistic regression
  - 2 decision trees
- Decision tree models generally showcase better predictability for non-functionality (True Positive), and slightly better for functionality (True Negative)

- Decision Tree (Partition): 101 splits
- Bootstrap Forest: >600 splits

The table below compares the results from the different models generated. The decision tree and bootstrap forest models has a higher prediction of true positives of 0.75 – 0.79. The decision tree, however, has higher Type I error of 0.34. Between the decision tree and bootstrap forest, the latter seems to provide a better decision than decision tree. However, the bootstrap forest's non-functionality prediction for the training set, 0.79, is much higher than the validation and training sets of 0.76. This may be an indication of overfitting in the bootstrap forest model.

Model Performance

	Method	TP	FN	TN	FP	Sensitivity	Specificity	Precision	Accuracy	F1	MCC	Profit
Training	Predictor Fit Nominal Logistic	8790	3682	13328	6944	0.7048	0.6575	0.5587	0.6755	0.6233	0.3521	-0.242
	Predictor Stepwise (Forward)	8715	3757	13325	6947	0.6988	0.6573	0.5564	0.6731	0.6195	0.3462	-0.245
	Predictor Stepwise (Backward)	8752	3720	13307	6965	0.7017	0.6564	0.5568	0.6737	0.621	0.3481	-0.244
	Predictor Partition	9404	3068	13416	6856	0.7540	0.6618	0.5784	0.6969	0.6546	0.4038	-0.222
	Predictor Bootstrap Forest	9857	2615	14025	6247	0.7903	0.6918	0.6121	0.7294	0.6899	0.4684	-0.197
	Method	**TP**	**FN**	**TN**	**FP**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**	**F1**	**MCC**	**Profit**
Validation	Predictor Fit Nominal Logistic	6534	2821	10049	5154	0.6985	0.6610	0.559	0.6753	0.621	0.3495	-0.244
	Predictor Stepwise (Forward)	6496	2859	10011	5192	0.6944	0.6585	0.5558	0.6722	0.6174	0.3431	-0.246
	Predictor Stepwise (Backward)	6537	2818	10015	5188	0.6988	0.6588	0.5575	0.674	0.6202	0.3476	-0.244
	Predictor Partition	6943	2412	10014	5189	0.7422	0.6587	0.5723	0.6905	0.6463	0.3894	-0.228
	Predictor Bootstrap Forest	7083	2272	10225	4978	0.7571	0.6726	0.5873	0.7048	0.6615	0.4174	-0.217
	Method	**TP**	**FN**	**TN**	**FP**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**	**F1**	**MCC**	**Profit**
Testing	Predictor Fit Nominal Logistic	6576	2778	10107	5097	0.7030	0.6648	0.5634	0.6793	0.6255	0.3576	-0.24
	Predictor Stepwise (Forward)	6556	2798	10113	5091	0.7009	0.6652	0.5629	0.6788	0.6244	0.356	-0.241
	Predictor Stepwise (Backward)	6592	2762	10120	5084	0.7047	0.6656	0.5646	0.6805	0.6269	0.3601	-0.239
	Predictor Partition	6928	2426	10091	5113	0.7406	0.6637	0.5754	0.693	0.6476	0.3928	-0.226
	Predictor Bootstrap Forest	7122	2232	10328	4876	0.7614	0.6793	0.5936	0.7106	0.6671	0.4281	-0.213

- Decision tree models showcase higher sensitivity, and accuracy than logistic regression
- Specificity and precision on par for some sample sets

In general, based on the performance metrics, the decision tree and bootstrap forest models have better sensitivity, specificity, precision, and accuracy than the logistic regression models.

Model Variables

	Variable that contributes to non-functionality prediction				
Variables	**Log. Reg.**	**Stepwise (Forward)**	**Stepwise (Backward)**	**Decision Tree**	**Bootstrap Forest**
Subjective quality of water	No	No		Mixed	
Local population within 1km	<10000	<10000	<10000	<1000	>10000
Payment required	No	No		No	Yes
Crucialness of system	Higher	Higher		Higher	Lower
Population served	<10000	<10000			
Water sources	Borehole	Borehole		Borehole	Others
Distance to town		Higher			
Urban area		True	True	True	
Administrative Area	Various	Various	Various	Various	Various

- Bootstrap Forest tend to contradict the other models

Various: Abia, Akwa Ibom, Cross River, Gombe, Kogi, Ondo, Teraba

The variables that contribute to the model predictions are shown in the table below. Most of the models show that a low local population, free water system, higher crucialness, urban areas and various administrative areas (Abia, Akwa Ibom, Cross River, Gombe, Kogi, Ondo, Teraba) tend to have higher contribution to non-functionality. The bootstrap forest model tends to contradict the parameters for the other models.

Higher crucialness and urban areas tend to signify higher usage of water. Low local population may also suggest lower population density, and hence the water system may be more spread out. This may result in stagnation in certain parts of the system, resulting in higher risk of bacterial growth. A water system with no payment required also means funds may need to be acquired elsewhere for maintenance of the system, which may be more challenging than a system that requires payment.

Conclusion

- Based on model performance, complexity, and variables, most suitable model seems to be Decision Tree
- Variables that affect the model include:
  - Local population within 1km (<1000)
  - Payment system (No payment)
  - Crucialness of system (High crucialness)
  - Water sources (Borehole)
  - Whether the system is in an urban area (Yes)
  - Various administrative area

In conclusion, the analysis has churned out several models to predict the non-functionality of the water systems in Nigeria. Out of all the models, the Decision Tree was deemed best in terms of performance, complexity, and variable sensibility.

The variables that affect the model are the local population within 1km, payment system, crucialness of system, water sources, whether the system is in an urban area and various administrative area.

Recommendations

1. Priority to be put into urban areas with low population but high crucialness of system
2. Suggest introducing fair pricing for usage of water system
3. Subjective quality of water needs to be accompanied by scientific quantitative analysis
4. Further investigation into administrative areas required

a. Priority to be put into urban areas with low population but high crucialness of system

Based on the model, these areas seem to signal higher non-functionality in water systems. By prioritising the rehabilitation or maintenance of these systems, any reliability issues can be addressed in a timely manner. Hence, ensuring that clean and safe water can be delivered in a reliable and sustainable manner throughout the country.

a. Suggest introducing fair pricing for usage of water system

A water system that does not require pricing may result in misuse of systems, and hence increase its risk of non-functionality. However, a fair pricing needs to be introduced and govern to ensure that the supplier's do not exploit the market for an essential commodity.

a. Subjective quality of water needs to be accompanied by scientific

quantitative analysis

The subjective quality of water alone may be insufficient to determine the usability of the water as it only covers colour, odour, and taste. Other components that affect the quality of the water such as fecal matter, pH, microorganism and bacterial growth cannot be assessed through this.

a. Further investigation into administrative areas required

Various administrative areas (Abia, Akwa Ibom, Cross River, Gombe, Kogi, Ondo, Teraba) seem to have higher contribution to non-functionality. Further analysis may need to be carried out for these areas to identify the root cause (such as low funding etc.) and apply the corrective measures to ensure that the area is accessible to safe water.