Nigeria Water Quality Prediction by Machine Learning

Ranice TAN Hui Qi

Contents

1. OVERVIEW 2

2. OBJECTIVE 2

3. DATA 2

3.1. DATA USED 2
3.2. DATA PREPARATION 2
3.3. DATA QUALITY 3
a. Reviewing the data and metadata 3
b. Checking for missing data 3
c. Univariate data analysis 4
d. Bivariate data analysis 4
e. Data Sampling 5

4. DATA ANALYSIS 6

4.1. CALIBRATING BASE MODEL 6
4.2. NOMINAL LOGISTIC REGRESSION 7
4.3. STEPWISE LOGISTIC REGRESSION (FORWARD) 8
4.4. STEPWISE LOGISTIC REGRESSION (BACKWARD) 10
4.5. DECISION TREE 11
4.6. BOOTSTRAP FOREST 12
4.7. MODEL COMPARISON 13

5. CONCLUSION 16

6. RECOMMENDATION 16

7. APPENDIX 17

7.1. APPENDIX A: FIELDS IN ORIGINAL DATASET 17
7.2. APPENDIX B: DATA PREPARATION CHANGE LOG 19

1. Overview

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.

2. Objective

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.

3. Data

3.1. Data Used

The dataset is obtained from the Water Point Data Exchange (WPdx) Data Repository[1]. 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. The data for 'Nigeria' was filtered and exported as follows:

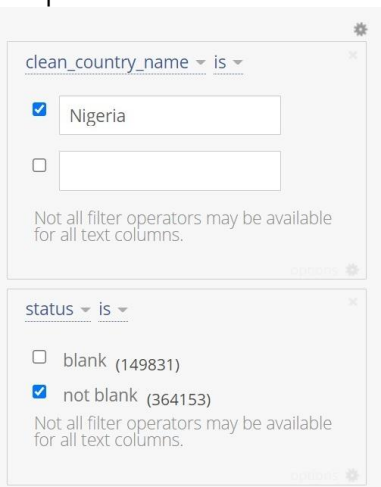


Figure 1 Filter criterion for WPdx Data Repository

3.2. Data Preparation

The dataset was imported into JMP Pro to ensure all fields are filled, and all columns are appropriately formatted. The fields in the original dataset can be found in 17. A preliminary inspection of the data summary statistics and distribution was also conducted. The data preparation log is accessible in 19.

[1] https://data.waterpointdata.org/dataset/Water-Point-Data-Exchange-Plus-WPdx-/eqje-vguj/data

3.3. Data Quality

a. Reviewing the data and metadata

The raw data was extracted and reviewed against the WPdx data standard[2]. The list of fields in the dataset can be found in 7.1. The metadata was first explored to ensure all data is in the correct form. Next, data that are not updated were removed from the dataset. Finally, the data that are irrelevant or redundant to the purpose of the analysis were removed. The amendments made can be found in 1-36 in the Data Change Log (19).

b. Checking for missing data

Column viewer was then used to generate the summary statistics as shown below:

Summary Statistics

14 Columns Clear Select Distribution

Columns	N	N Missing	Min	Max	Mean	Std Dev
lat_deg	88617	0	4.3018117	13.865675	9.55923455189298	2.48445275901823
lon_deg	88617	0	2.707441	13.502203	7.4991706845436	2.2142247149987
report_date	88617	0	01/01/2010 12:00:00 am	02/22/2016 12:00:00 am	05/05/2015 7:46:03.5 am	:108:18:41
install_year	48833	39784	1904	2045	2007.07496979502	8.19102063009088
distance_to_primary_road	88617	0	0.0144613568133352	82056.9295146987	10023.8509087787	11193.3810836815
distance_to_secondary_road	88617	0	0.00463002014735481	94773.0323900959	8030.13214539517	10285.8740095833
distance_to_tertiary_road	88617	0	0.00433615499544107	58873.9896748268	3919.68846442285	5929.4584043526
distance_to_city	88617	0	12.1166340884705	216727.25388027	49649.9466036921	36787.6748951094
distance_to_town	88617	0	10.0006931278202	97758.6982663442	21571.524942144	18220.3623861744
rehab_priority	35026	53591	0	389230	1429.70847370525	4737.58074714701
served_population	88343	274	0	389230	1150.31387885854	3616.27533862284
local_population_1km	88343	274	0	389230	3755.90955706734	7100.59649832738
crucialness	83001	5616	0.0000401412973667309	1.00016616816218	0.375731469759287	0.32491324227427
pressure	82174	6443	0.001	466.182	1.84809056392533	4.8439105368456

Figure 2 Summary Statistics of Continuous Variables

Summary Statistics

21 Columns Clear Select Distribution

Columns	N	N Missing	N Categories
water_source_clean	88617	0	7
water_source_category	88617	0	3
water_tech_clean	87407	1224	5
water_tech_category	87407	1224	3
facility_type	88617	0	2
clean_country_name	88617	0	1
clean_adm1	88617	0	37
clean_adm2	88617	0	746
clean_adm3	0	89369	0
installer	0	89369	0
rehab_year	0	89369	0
rehabilitator	0	89369	0
management_clean	61546	27352	7
pay	88617	0	42
status	88617	0	881
fecal_coliform_presence	0	89369	0
fecal_coliform_value	0	89369	0
subjective_quality	88617	0	4
usage_capacity	87738	892	3
is_urban	88617	0	2

Figure 3 Summary Statistics of Nominal Variables

Firstly, columns with relatively large amount of missing data (in red) were excluded (> 25000 rows) to avoid inaccurate model due to lack of data. Next, nominal variables with large number of categories were either removed (in red) or recoded (in green) based on the importance of the variables. Lastly, useful variables with rows containing

[2] https://www.waterpointdata.org/wp-content/uploads/2021/04/WPDx_Data_Standard.pdf

missing data (in yellow) were excluded to prevent interfering with the model results. A summary of the changes made can be found in 37-57 in the Data Change Log (19).

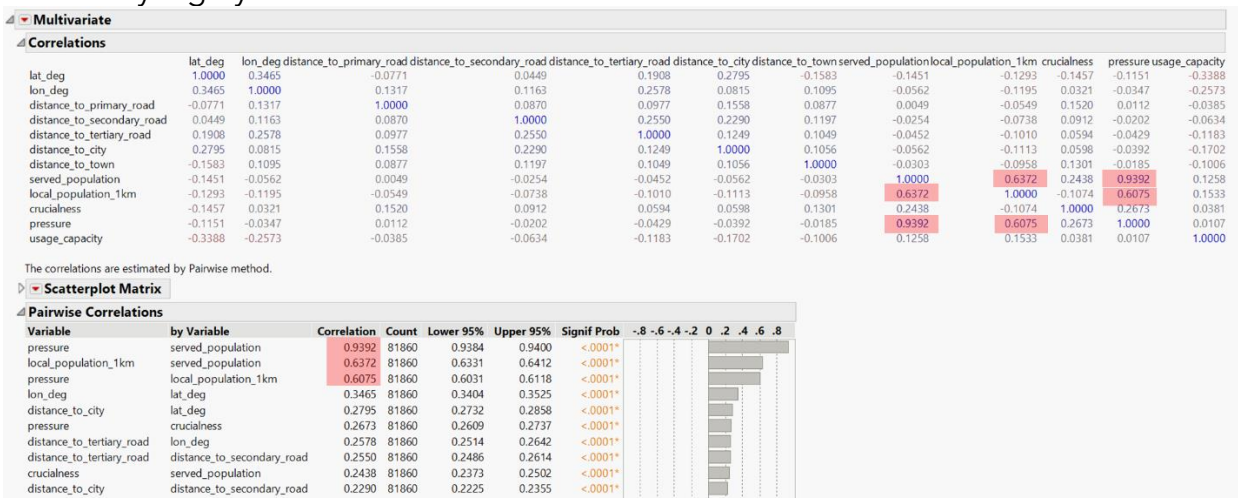
c. Univariate data analysis

The distribution of the remaining variables was explored and those with relatively uneven distribution were identified below. Variables with uneven distribution were recoded in 58-61 of the Data Change Log (19).

Figure 4 Uneven Distribution of variables

d. Bivariate data analysis

A pairwise correlation for continuous variables was then carried out to identify highly correlated variables that can be interchanged. Based on the results below, the variables pressure, served_population and local_population_1km seems to be relatively highly correlated with one another.



Multivariate

Correlations

	lat_deg	lon_deg	distance_to_primary_road	distance_to_secondary_road	distance_to_tertiary_road	distance_to_city	distance_to_town	served_population	local_population_1km	crucialness	pressure	usage_capacity
lat_deg	1.0000	0.3465	-0.0771	0.0449	0.1908	0.2795	-0.1583	-0.1451	-0.1293	-0.1457	-0.1151	-0.3388
lon_deg	0.3465	1.0000	0.1317	0.1163	0.2578	0.0815	0.1095	-0.0562	-0.1195	0.0321	-0.0347	-0.2573
distance_to_primary_road	-0.0771	0.1317	1.0000	0.0870	0.0977	0.1558	0.0877	0.0049	-0.0549	0.1520	0.0112	-0.0385
distance_to_secondary_road	0.0449	0.1163	0.0870	1.0000	0.2550	0.2290	0.1197	-0.0254	-0.0738	0.0912	-0.0202	-0.0634
distance_to_tertiary_road	0.1908	0.2578	0.0977	0.2550	1.0000	0.1249	0.1049	-0.0452	-0.1010	0.0594	-0.0429	-0.1183
distance_to_city	0.2795	0.0815	0.1558	0.2290	0.1249	1.0000	0.1056	-0.0562	-0.1113	0.0598	-0.0392	-0.1702
distance_to_town	-0.1583	0.1095	0.0877	0.1197	0.1049	0.1056	1.0000	-0.0303	-0.0958	0.1301	-0.0185	-0.1006
served_population	-0.1451	-0.0562	0.0049	-0.0254	-0.0452	-0.0562	-0.0303	1.0000	0.6372	0.2438	0.9392	0.1258
local_population_1km	-0.1293	-0.1195	-0.0549	-0.0738	-0.1010	-0.1113	-0.0958	0.6372	1.0000	-0.1074	0.6075	0.1533
crucialness	-0.1457	0.0321	0.1520	0.0912	0.0594	0.0598	0.1301	0.2438	-0.1074	1.0000	0.2673	0.0381
pressure	-0.1151	-0.0347	0.0112	-0.0202	-0.0429	-0.0392	-0.0185	0.9392	0.6075	0.2673	1.0000	0.0107
usage_capacity	-0.3388	-0.2573	-0.0385	-0.0634	-0.1183	-0.1702	-0.1006	0.1258	0.1533	0.0381	0.0107	1.0000

The correlations are estimated by Pairwise method.

Scatterplot Matrix

Pairwise Correlations

Variable	by Variable	Correlation	Count	Lower 95%	Upper 95%	Signif Prob
pressure	served_population	0.9392	81860	0.9384	0.9400	<.0001*
local_population_1km	served_population	0.6372	81860	0.6331	0.6412	<.0001*
pressure	local_population_1km	0.6075	81860	0.6031	0.6118	<.0001*
lon_deg	lat_deg	0.3465	81860	0.3404	0.3525	<.0001*
distance_to_city	lat_deg	0.2795	81860	0.2732	0.2858	<.0001*
pressure	crucialness	0.2673	81860	0.2609	0.2737	<.0001*
distance_to_tertiary_road	lon_deg	0.2578	81860	0.2514	0.2642	<.0001*
distance_to_tertiary_road	distance_to_secondary_road	0.2550	81860	0.2486	0.2614	<.0001*
crucialness	served_population	0.2438	81860	0.2373	0.2502	<.0001*
distance_to_city	distance_to_secondary_road	0.2290	81860	0.2225	0.2355	<.0001*

Figure 5 Pairwise Correlations of continuous variables

A principal component analysis was then conducted to identify how many variables can be excluded. Based on the eigenvalue table, the first 3 principal components have an eigenvalue of greater than 1.

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	2.6177	23.797		23.797
2	1.8173	16.521		40.319
3	1.3779	12.527		52.845
4	0.9762	8.874		61.719
5	0.9596	8.724		70.443
6	0.8763	7.966		78.409
7	0.8543	7.766		86.176
8	0.6784	6.168		92.343
9	0.4423	4.021		96.364
10	0.3406	3.096		99.461
11	0.0593	0.539		100.000

Figure 6 Eigenvalues of pairwise correlation variables

Finally, variable pressure was excluded and the analysis proceeded with 10 variables, accounting for 99% of the variance.

e. Data Sampling

A validation column was added to split the data into training, validation and test set, stratified by the STATUS recoded column.

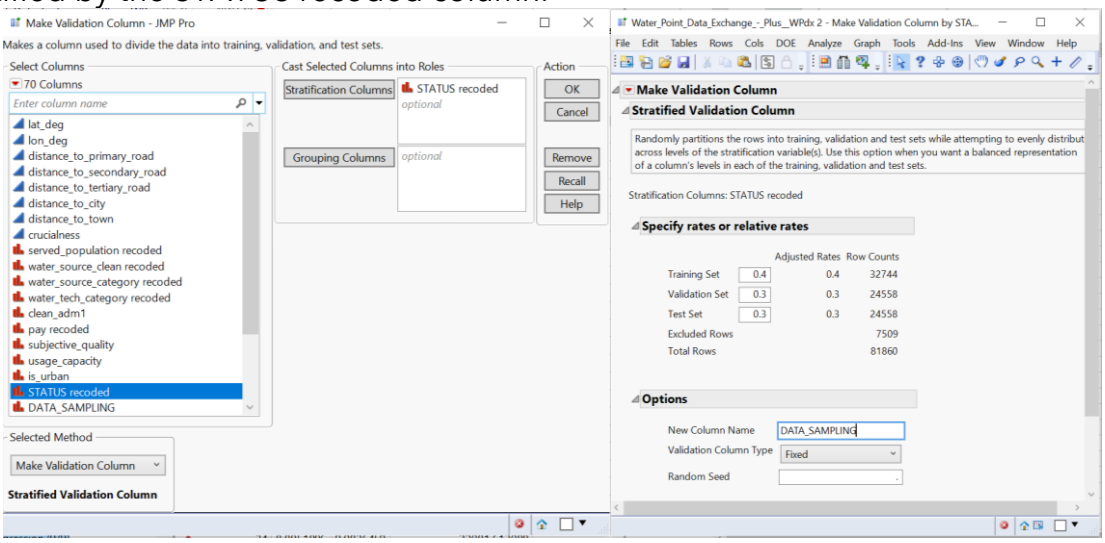


Figure 7 Making Validation Column

4. Data Analysis

4.1. Calibrating Base Model

In the dataset, there are a total of 20 predictors. 8 of them are continuous predictors and the other 12 are categorial predictors. Among the categorical predictors, only 3 is binary and the remaining 9 have more than two classes. The base nominal logistic model was generated as follows:

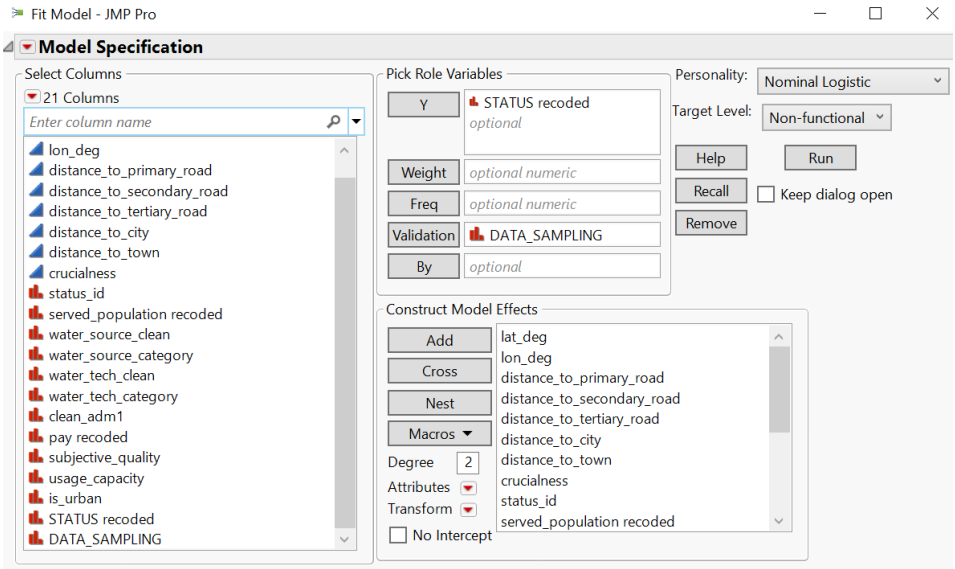


Figure 8 Model specification for base model

By interpretating the model through the parameter estimates, the status_id is marked as an unstable estimate, and the water_sources and water_techs are marked as biased or zeroed. The binary status_id variable was then excluded from the model, the water_source_clean, water_source_category and water_tech_category was recoded into binary predictors. water_tech_clean was removed as if recoded into binary, it would be a duplicate of water_tech_category.

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Biased	-5.7722e-9	1652264.8	0.00	1.0000
lat_deg		1.6704e-13	72429.936	0.00	1.0000
lon_deg		1.1454e-13	74851.357	0.00	1.0000
distance_to_primary_road		2.0666e-19	3.2590273	0.00	1.0000
distance_to_secondary_road		2.017e-19	3.8027027	0.00	1.0000
distance_to_tertiary_road		-1.53e-18	6.4373221	0.00	1.0000
distance_to_city		6.325e-19	1.2224955	0.00	1.0000
distance_to_town		5.0114e-19	2.587149	0.00	1.0000
crucialness		3.5953e-14	124411.63	0.00	1.0000
status_id[No]	Unstable	31.2028948	36816.845	0.00	0.9993
served_population recoded[> 100000]		-5.02e-13	2008674.4	0.00	1.0000
served_population recoded[1 - 1000]		1.6897e-13	676560.34	0.00	1.0000
served_population recoded[1001 - 10000]		1.725e-13	675170.21	0.00	1.0000
water_source_clean[Borehole]	Biased	-6.937e-15	175919.61	0.00	1.0000
water_source_clean[Protected Shallow Well]	Biased	-6.842e-15	215360.23	0.00	1.0000
water_source_clean[Protected Spring]	Zeroed	0	0	.	.
water_source_category[Spring]	Zeroed	0	0	.	.
water_tech_clean[Hand Pump]	Biased	9.7402e-13	1272044.2	0.00	1.0000
water_tech_clean[Mechanized Pump]	Biased	-2.95e-13	1193406.3	0.00	1.0000
water_tech_clean[Mechanized Pump - Diesel]	Biased	-1.748e-12	2936572.6	0.00	1.0000
water_tech_clean[Mechanized Pump - Solar]	Biased	-3.087e-13	1245664	0.00	1.0000
water_tech_category[Hand Pump]	Zeroed	0	0	.	.
water_tech_category[Mechanized Pump]	Zeroed	0	0	.	.
clean_adm1[Abia]		5.537e-13	389781.24	0.00	1.0000
clean_adm1[Adamawa]		-7.953e-13	693517.59	0.00	1.0000
clean_adm1[Akwa Ibom]		5.5491e-13	355069.62	0.00	1.0000
clean_adm1[Anambra]		4.4838e-13	468434.43	0.00	1.0000

Figure 9 Fit Details and Parameter Estimates of Base Model

The final dataset consists of a total of 17 predictors - 8 which are continuous and 9 which are categorial predictors. Among the 9 categorical predictors, 4 are binary and the remaining 5 have more than two classes.

4.2. Nominal Logistic Regression

Using the revised data table from calibration, the result below was achieved. The model has a misclassification rate of 0.3, a non-functional prediction rate of 0.47, and a functional prediction rate of 0.83. The false positive error is 0.52 and false negative error is 0.17.

Nominal Logistic Fit for STATUS recoded

Whole Model Test

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1315	0.1339	0.1330	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2181	0.2217	0.2204	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5771	0.5755	0.5761	∑ -Log(ρ[j])/n
RASE	0.4436	0.4430	0.4430	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3937	0.3924	0.3924	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.3054	0.3044	0.3045	∑ (ρ[j]≠ρMax)/n
N	32744	24558	24558	n

Lack Of Fit

Parameter Estimates

Effect Likelihood Ratio Tests

Confusion Matrix

Training

Actual STATUS recoded	Predicted Count Non-functional	Functional
Non-functional	5969	6503
Functional	3498	16774

Actual STATUS recoded	Predicted Rate Non-functional	Functional
Non-functional	0.479	0.521
Functional	0.173	0.827

Validation

Actual STATUS recoded	Predicted Count Non-functional	Functional
Non-functional	4438	4917
Functional	2559	12644

Actual STATUS recoded	Predicted Rate Non-functional	Functional
Non-functional	0.474	0.526
Functional	0.168	0.832

Test

Actual STATUS recoded	Predicted Count Non-functional	Functional
Non-functional	4428	4926
Functional	2553	12651

Actual STATUS recoded	Predicted Rate Non-functional	Functional
Non-functional	0.473	0.527
Functional	0.168	0.832

Figure 10 Results for Nominal Logistic Regression

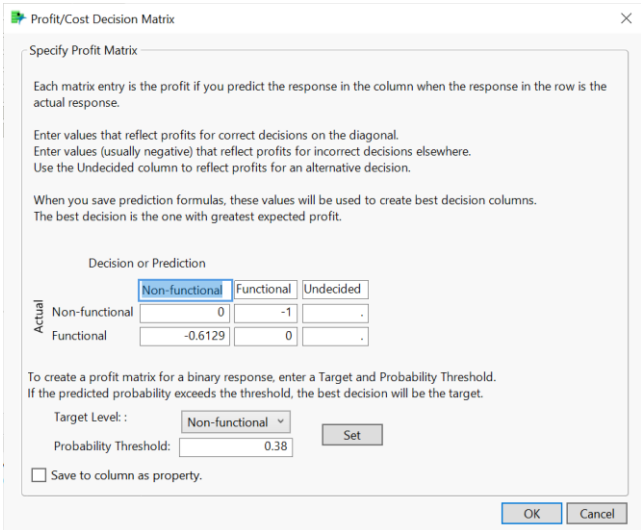


Figure 11 Specify probability threshold for non-functional level

From the effect summary, the variables that influence the model most are administrative area, quality of water, local population within 1km of the water system, payment system, crucialness of system, population served and water source.

Effect Summary

Source	LogWorth	PValue
clean_adm1	172.907	0.00000
subjective_quality	137.291	0.00000
local_population_1km recoded	109.566	0.00000
pay recoded	89.266	0.00000
crucialness	71.333	0.00000
served_population recoded	40.587	0.00000
water_source_clean recoded	35.093	0.00000
lat_deg	5.520	0.00000
is_urban	5.257	0.00001
lon_deg	4.635	0.00002
distance_to_town	1.577	0.02649
distance_to_secondary_road	0.664	0.21672
distance_to_primary_road	0.341	0.45555
distance_to_city	0.291	0.51170
usage_capacity	0.254	0.55708
water_tech_clean recoded	0.173	0.67076
distance_to_tertiary_road	0.022	0.95035

Figure 12 Effect Summary of Nominal Logistic Regression Model

From the formula of the model below, the non-functional water systems have population within 1km and served population ≤ 10000. The sources also tend to come from boreholes, do not require payment, have unacceptable quality of water, and a low usage capacity of 250.

Lower population means lower usage of water, which may have more stagnant pipelines, resulting in an increased risk of biofilm-associated bacteria, or spread of Legionella. Hence, resulting in unacceptable quality of water. Without payment for usage of the water system, more funds will need to be obtained to carry out preventive maintenance.

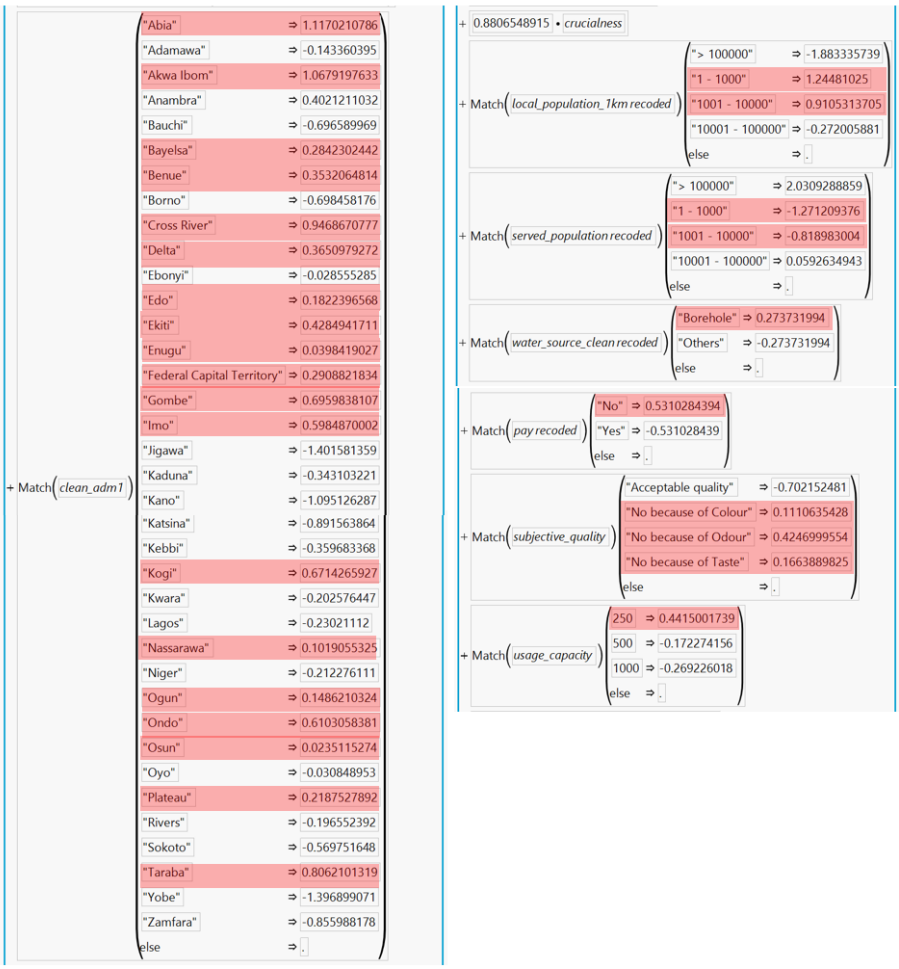


Figure 13 Formulas of variables that impact the reliability of water system

4.3. Stepwise Logistic Regression (Forward)

A forward stepwise logistic regression was also carried out as follows. Minimum BIC was selected as the stopping rule due to the large dataset (~80000 rows) acquired.

Stepwise Regression Control

Stopping Rule: Minimum BIC

Direction: Forward

Rules: Combine

Enter All | Make Model | Remove All | Run Model

Go | Stop | Step

-LogLikelihood	p	RSquare	AICc	BIC	RSquare Validation	Avg Log Error Validation	RSquare Test	Avg Log Error Test
18971.986	18	0.1281	37980	38131.1	0.1308	0.577595	0.1315	0.577141

Current Estimates

Step History

Step	Parameter	Action	L-R ChiSquare	"Sig Prob"	Entry ChiSquare	Entry "Sig Prob"	RSquare	p	AICc	BIC	RSquare Validation
1	clean_adm1{Akwa Ibom&Abia&Kogi&Bayelsa&Ondo&Cross River&Delta&Edo&Imo&Ogun&Gomb...	Entered	2246.862	0.0000	2143.72	0	0.0516	2	41273.9	41290.7	0.0560
2	crucialness	Entered	1273.27	0.0000	1282.87	6e-281	0.0809	3	40002.7	40027.8	0.0827
3	subjective_quality{No because of Odour&No because of Colour&No because of Taste-Acceptable q...	Entered	607.3388	0.0000	615.58	7e-136	0.0948	4	39397.3	39430.9	0.0955
4	pay recoded{No-Yes}	Entered	371.7533	0.0000	345.051	5.1e-77	0.1034	5	39027.6	39069.5	0.1029
5	local_population_1km recoded{1 - 1000&1001 - 10000-10001 - 100000&> 100000}	Entered	207.3226	0.0000	198.269	5e-45	0.1082	6	38822.2	38872.6	0.1088
6	clean_adm1{Akwa Ibom&Abia&Kogi&Bayelsa&Ondo&Cross River&Delta&Edo-Imo&Ogun&Gombe...	Entered	135.3089	0.0000	135.197	3e-31	0.1113	7	38688.9	38747.7	0.1147
7	water_source_clean recoded{Borehole-Others}	Entered	120.3641	0.0000	116.701	3.3e-27	0.1140	8	38570.6	38637.7	0.1170
8	clean_adm1{Zamfara&Adamawa&Katsina&Rivers&Bauchi&Kano-Borno&Jigawa&Yobe}	Entered	119.0734	0.0000	115.474	6.2e-27	0.1168	9	38453.5	38529.1	0.1200
9	served_population recoded{> 100000&10001 - 100000&1001 - 10000-1 - 1000}	Entered	83.40008	0.0000	83.8532	5.3e-20	0.1187	10	38372.1	38456.1	0.1213
10	local_population_1km recoded{1 - 1000-1001 - 10000}	Entered	99.23582	0.0000	99.6016	1.9e-23	0.1210	11	38274.9	38367.2	0.1234
11	distance_to_town	Entered	68.26638	0.0000	67.8556	1.8e-16	0.1225	12	38208.6	38309.4	0.1248
12	served_population recoded{> 100000&10001 - 100000-1001 - 10000}	Entered	53.86046	0.0000	54.4696	1.6e-13	0.1238	13	38156.7	38265.9	0.1256
13	clean_adm1{Zamfara&Adamawa&Katsina&Rivers-Bauchi&Kano}	Entered	48.17534	0.0000	48.4826	3.3e-12	0.1249	14	38110.6	38228.1	0.1275
14	clean_adm1{Imo&Ogun&Gombe&Ekiti&Kebbi&Federal Capital Territory&Benue&Anambra&Kwara...	Entered	42.42542	0.0000	42.3143	7.8e-11	0.1259	15	38070.1	38196.1	0.1290
15	clean_adm1{Nassarawa&Oyo&Niger-Ebonyi&Kaduna&Lagos}	Entered	37.18721	0.0000	37.1873	1.07e-9	0.1267	16	38035	38169.3	0.1299
16	is_urban{False-True}	Entered	33.12806	0.0000	33.2669	8.03e-9	0.1275	17	38003.8	38146.6	0.1304
17	clean_adm1{Imo&Ogun&Gombe&Ekiti&Kebbi&Federal Capital Territory-Benue&Anambra&Kwara&...	Entered	25.84411	0.0000	25.8506	3.69e-7	0.1281	18	37980	38131.1	0.1308
18	clean_adm1{Benue&Anambra&Kwara&Plateau-Sokoto&Taraba&Enugu&Osun}	Entered	10.21655	0.0014	10.2165	0.00139	0.1283	19	37971.8	38131.3	0.1309
19	clean_adm1{Kogi&Bayelsa&Ondo-Cross River&Delta&Edo}	Entered	16.76506	0.0000	13.765	0.00103	0.1287	21	37959	38135.3	0.1315
20	subjective_quality{No because of Odour-No because of Colour}	Entered	10.22823	0.0014	10.0225	0.00666	0.1289	23	37952.8	38145.9	0.1318

Figure 14 Stepwise (Forward) Regression Control

The model has a misclassification rate of 0.31, a non-functional prediction rate of 0.47, and a functional prediction rate of 0.83. The false positive error is 0.53 and false negative error is 0.17.

Stepwise (Forward)

Fit Details

Measure	Training	Validation	Test	Definition
Generalized RSquare	0.2128	0.2170	0.2180	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5794	0.5776	0.5771	∑ -Log(p[j])/n
RASE	0.4447	0.4440	0.4434	√ ∑(y[j]-p[j])²/n
Mean Abs Dev	0.3956	0.3942	0.3937	∑ |y[j]-p[j]|/n
Misclassification Rate	0.3067	0.3083	0.3075	∑ (p[j]≠pMax)/n
N	32744	24558	24558	n

Lack Of Fit

Parameter Estimates

Effect Likelihood Ratio Tests

Confusion Matrix

Training

Actual STATUS recoded	Predicted Count Non-functional	Functional
Non-functional	5874	6598
Functional	3443	16829

Actual STATUS recoded	Predicted Rate Non-functional	Functional
Non-functional	0.471	0.529
Functional	0.170	0.830

Validation

Actual STATUS recoded	Predicted Count Non-functional	Functional
Non-functional	4368	4987
Functional	2585	12618

Actual STATUS recoded	Predicted Rate Non-functional	Functional
Non-functional	0.467	0.533
Functional	0.170	0.830

Test

Actual STATUS recoded	Predicted Count Non-functional	Functional
Non-functional	4354	5000
Functional	2552	12652

Actual STATUS recoded	Predicted Rate Non-functional	Functional
Non-functional	0.465	0.535
Functional	0.168	0.832

Figure 15 Results for Stepwise Logistic Regression (Forward)

From the step history, the variables that influence the model most are administrative area, crucialness of system, quality of water, payment system, local population within 1km of the water system, water source and population served, distance to town and whether the system is in an urban area.

The variables are like the nominal logistic regression except 2 variables – distance to town and whether the system is in an urban area. For the variables that are same, they are also directionally similar for predicting non-functionality. Water systems in urban areas have greater chance of being non-functional. This may be due to the higher utilisation rate of water in urban areas compared to rural areas. Hence, higher priority may need to be put for maintenance or rehabilitation of crucial water systems in urban areas.

```
+ Match( is_urban )
    "False" ⇒ -0.103804146
    "True"  ⇒ 0.1038041457
    else    ⇒ .
```

Figure 16 Urban formula in Stepwise Logistic Regression (Forward)

4.4. Stepwise Logistic Regression (Backward)

A backward stepwise logistic regression was also carried out as follows. Minimum BIC was selected as the stopping rule due to the large dataset (~80000 rows) acquired.

Stepwise Regression Control

Stopping Rule: Minimum BIC
Direction: Backward
Rules: Combine

Enter All | Make Model | Remove All | Run Model | Go | Stop | Step

-LogLikelihood	p	RSquare	AICc	BIC	RSquare Validation	Avg Log Error Validation	RSquare Test	Avg Log Error Test
18965.114	20	0.1284	37970.3	38138.2	0.1311	0.577391	0.1319	0.576873

Current Estimates

Step History

Step	Parameter	Action	L-R ChiSquare	"Sig Prob"	Entry ChiSquare	Entry "Sig Prob"	RSquare	p	AICc	BIC	RSquare Validation
1	clean_adm1{Jigawa-Yobe}	Removed	0.001015	0.9746	0.00102	0.97457	0.1315	59	37912.5	38407.7	0.1339
2	distance_to_tertiary_road	Removed	0.003739	0.9512	0.00374	0.95124	0.1315	58	37910.5	38397.3	0.1339
3	clean_adm1{Enugu-Osun}	Removed	0.006288	0.9368	0.00629	0.93679	0.1315	57	37908.5	38386.9	0.1339
4	clean_adm1{Akwa Ibom-Abia}	Removed	0.122491	0.7263	0.12242	0.72643	0.1315	56	37906.6	38376.6	0.1339
5	water_tech_clean recoded{Others-Hand Pump}	Removed	0.183531	0.6684	0.18533	0.66683	0.1315	55	37904.8	38366.4	0.1339
6	distance_to_city	Removed	0.44828	0.5032	0.44823	0.50318	0.1315	54	37903.3	38356.5	0.1339
7	clean_adm1{Nassarawa-Oyo}	Removed	0.646648	0.4213	0.64686	0.42124	0.1315	53	37901.9	38346.7	0.1338
8	clean_adm1{Kogi&Bayelsa-Ondo}	Removed	1.792962	0.1806	1.8072	0.40511	0.1314	51	37899.7	38327.7	0.1338
9	distance_to_secondary_road	Removed	0.897968	0.3433	0.89818	0.34327	0.1314	50	37898.6	38318.2	0.1337
10	distance_to_primary_road	Removed	1.018449	0.3129	1.01795	0.31301	0.1314	49	37897.6	38308.9	0.1337
11	usage_capacity{500-250}	Removed	1.446489	0.2291	1.60271	0.20552	0.1313	48	37897	38299.9	0.1337
12	served_population recoded{> 100000-10001 - 100000}	Removed	2.0156	0.1557	1.65195	0.19869	0.1313	47	37897	38291.5	0.1337
13	local_population_1km recoded{10001 - 100000-> 100000}	Removed	1.787422	0.1812	1.64237	0.2	0.1313	46	37896.8	38282.9	0.1336
14	clean_adm1{Ogun-Gombe}	Removed	3.428748	0.0641	3.42795	0.0641	0.1312	45	37898.2	38276	0.1333
15	clean_adm1{Imo-Ogun&Gombe}	Removed	3.393116	0.0655	3.39164	0.06553	0.1311	44	37899.6	38269	0.1335
16	clean_adm1{Zamfara-Adamawa&Katsina}	Removed	6.487007	0.0109	6.73877	0.03441	0.1310	42	37902.1	38254.6	0.1336
17	clean_adm1{Ebonyi-Kaduna}	Removed	4.143382	0.0418	4.16482	0.04127	0.1309	41	37904.2	38248.4	0.1334
18	distance_to_town	Removed	3.371149	0.0663	3.36828	0.06646	0.1308	40	37905.6	38241.4	0.1333
19	clean_adm1{Benue&Anambra&Kwara-Plateau}	Removed	8.543936	0.0035	8.53921	0.03609	0.1306	37	37908.1	38218.7	0.1334
20	clean_adm1{Borno-Jigawa&Yobe}	Removed	4.188844	0.0407	4.49326	0.03403	0.1305	36	37910.3	38212.5	0.1333
21	clean_adm1{Ebonyi&Kaduna-Lagos}	Removed	5.351937	0.0207	5.39823	0.02016	0.1304	35	37913.7	38207.5	0.1333
22	clean_adm1{Zamfara&Adamawa&Katsina-Rivers}	Removed	4.270344	0.0388	4.36315	0.03672	0.1303	34	37915.9	38201.3	0.1334
23	clean_adm1{Nassarawa&Oyo-Niger}	Removed	3.608376	0.0575	3.60005	0.05778	0.1302	33	37917.5	38194.6	0.1332
24	clean_adm1{Kebbi-Federal Capital Territory}	Removed	3.750513	0.0528	3.7633	0.05239	0.1301	32	37919.3	38187.9	0.1329
25	clean_adm1{Bauchi-Kano}	Removed	5.741828	0.0166	5.78601	0.01615	0.1300	31	37923	38183.3	0.1330
26	clean_adm1{Cross River-Delta}	Removed	5.602309	0.0179	5.64388	0.01752	0.1298	30	37926.6	38178.5	0.1328
27	subjective_quality{No because of Odour&No because of Colour-No because of Taste}	Removed	9.600926	0.0019	9.49039	0.00869	0.1296	28	37932.2	38167.3	0.1325
28	clean_adm1{Imo&Ogun&Gombe-Ekiti&Kebbi&Federal Capital Territory}	Removed	12.43555	0.0004	12.4033	0.00203	0.1293	26	37940.6	38158.9	0.1321
29	lat_deg	Removed	8.623444	0.0033	8.58492	0.00339	0.1291	25	37947.3	38157.1	0.1318
30	clean_adm1{Sokoto-Taraba}	Removed	5.367824	0.0205	5.36689	0.02052	0.1290	24	37950.6	38152.1	0.1316
31	usage_capacity{1000-500&250}	Removed	6.849294	0.0089	6.86195	0.00881	0.1289	23	37955.5	38148.6	0.1313
32	clean_adm1{Cross River&Delta-Edo}	Removed	6.0485	0.0139	6.04839	0.01392	0.1287	22	37959.5	38144.2	0.1315
33	clean_adm1{Sokoto&Taraba-Enugu&Osun}	Removed	7.379396	0.0066	7.37323	0.00662	0.1285	21	37964.9	38141.2	0.1314
34	lon_deg	Removed	7.35844	0.0067	7.35177	0.0067	0.1284	20	37970.3	38138.2	0.1311
35	clean_adm1{Kogi&Bayelsa&Ondo-Cross River&Delta&Edo}	Removed	17.74717	0.0000	17.7201	2.56e-5	0.1280	19	37986	38145.5	0.1305
36	clean_adm1{Akwa Ibom&Abia-Kogi&Bayelsa&Ondo&Cross River&Delta&Edo}	Removed	3.485413	0.0619	3.46389	0.06272	0.1279	18	37987.5	38138.6	0.1304
37	clean_adm1{Benue&Anambra&Kwara&Plateau-Sokoto&Taraba&Enugu&Osun}	Removed	21.32048	0.0000	21.3019	3.92e-6	0.1274	17	38006.8	38149.5	0.1301
38	clean_adm1{Zamfara&Adamawa&Katsina&Rivers-Bauchi&Kano}	Removed	36.67607	0.0000	36.7356	1.35e-9	0.1266	16	38041.5	38175.8	0.1285

Figure 17 Stepwise (Backward) Regression Control

The model has a misclassification rate of 0.31, a non-functional prediction rate of 0.47, and a functional prediction rate of 0.83. The false positive error is 0.53 and false negative error is 0.17.

Stepwise (Backward)

Fit Details

Measure	Training	Validation	Test	Definition
Generalized RSquare	0.2133	0.2175	0.2186	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5792	0.5774	0.5769	∑ -Log(ρ[j])/n
RASE	0.4445	0.4438	0.4433	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3954	0.3941	0.3935	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.3066	0.3085	0.3074	∑ (ρ[j]≠ρMax)/n
N	32744	24558	24558	n

Lack Of Fit

Parameter Estimates

Effect Likelihood Ratio Tests

Confusion Matrix

Training

Actual STATUS recoded	Predicted Count Non-functional	Functional
Non-functional	5901	6571
Functional	3468	16804

Actual STATUS recoded	Predicted Rate Non-functional	Functional
Non-functional	0.473	0.527
Functional	0.171	0.829

Validation

Actual STATUS recoded	Predicted Count Non-functional	Functional
Non-functional	4357	4998
Functional	2579	12624

Actual STATUS recoded	Predicted Rate Non-functional	Functional
Non-functional	0.466	0.534
Functional	0.170	0.830

Test

Actual STATUS recoded	Predicted Count Non-functional	Functional
Non-functional	4369	4985
Functional	2563	12641

Actual STATUS recoded	Predicted Rate Non-functional	Functional
Non-functional	0.467	0.533
Functional	0.169	0.831

Figure 18 Results for Stepwise Logistic Regression (Backward)

From the step history, the variables that influence the model most are local population within 1km, administrative area and whether the system is in an urban area. The variables are like some of the variables in the previous logistic regression. They are also directionally similar for predicting non-functionality.

4.5. Decision Tree

A partition model was carried out as follows:

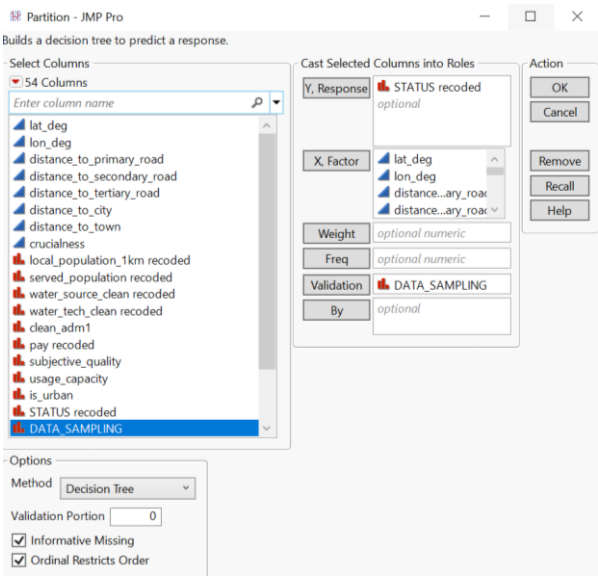


Figure 19 Partition Model Creation

The resulting model is a decision tree with 101 splits. The model has a misclassification rate of 0.29, a non-functional prediction rate of 0.61, and a functional prediction rate of 0.77. The false positive error is 0.39 and false negative error is 0.23.

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1816	0.1618	0.1586	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2916	0.2631	0.2585	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5438	0.5570	0.5591	∑ -Log(ρ[j])/n
RASE	0.4299	0.4357	0.4361	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3700	0.3748	0.3746	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2878	0.2956	0.2948	∑ (ρ[j]≠ρMax)/n
N	32744	24558	24558	n

Confusion Matrix

Training

Actual STATUS recoded	Predicted Count Functional	Non-functional
Functional	15591	4681
Non-functional	4742	7730

Actual STATUS recoded	Predicted Rate Functional	Non-functional
Functional	0.769	0.231
Non-functional	0.380	0.620

Validation

Actual STATUS recoded	Predicted Count Functional	Non-functional
Functional	11642	3561
Non-functional	3699	5656

Actual STATUS recoded	Predicted Rate Functional	Non-functional
Functional	0.766	0.234
Non-functional	0.395	0.605

Test

Actual STATUS recoded	Predicted Count Functional	Non-functional
Functional	11721	3483
Non-functional	3757	5597

Actual STATUS recoded	Predicted Rate Functional	Non-functional
Functional	0.771	0.229
Non-functional	0.402	0.598

Figure 20 Decision Tree Results

From the leaf report, the variables that contribute to non-functionality includes crucialness, subjective quality of the water, local population within 1km, payment method, water source and whether the system is in an urban area.

Water systems with higher crucialness > 0.09 and does not require payment show higher susceptibility to non-functionality. This may be because the systems have a higher usage, and without payment to help fund resources, it may be more difficult to carry out scheduled preventive maintenance. Water systems with lower crucialness (0.02<X<0.09) and local population of <1000 within 1km of an urban area, also have higher probability of non-functionality. Higher water usage may be expected in urban areas, but the low population cruciality and population rate may result in stagnant pipelines, which increases risk of bacteria growth, resulting in unacceptable quality of water. Both acceptable and non-acceptable subjective quality show high chance of non-functionality. This means that subjective quality alone is insufficient to determine the reliability system. This may be because it only covers colour, taste, and odour of water,

while other aspects such as microorganism and bacterial growth cannot be assessed through this.

Leaf Report

Response Prob

Leaf Label	Functional	.2 .4 .6 .8	Non-functional	.2 .4 .6 .8
^&crucialness<0.0912698413&crucialness>=0.0200892857&local_population_1km recoded(1 - 1000)&is_urban(True)^&clean_adm1(Ebonyi, ...	0.0174		0.9826	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)^&water_source_clean recoded(Borehole)&clean_ad...	0.0267		0.9733	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)^&water_source_clean recoded(Borehole)&clean_ad...	0.0346		0.9654	
^&crucialness>=0.0912698413&subjective_quality(No because of Odour, No because of Taste, No because of Colour)^&pay recoded(No)&l...	0.0739		0.9261	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)&clean_adm1(Abia, Kogi, Ondo, Gombe, Kebbi, Ogu...	0.0973		0.9027	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)&clean_adm1(Edo, Oyo, Benue, Anambra, Ekiti, Osu...	0.0995		0.9005	
clean_adm1(Akwa Ibom, Abia, Kogi, Ondo, Cross River, Bayelsa, Delta, Imo, Ogun, Kebbi, Ekiti, Edo, Anambra, Gombe, Benue, Sokoto, Plateau, ...	0.1220		0.8780	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)^&water_source_clean recoded(Borehole)&clean_ad...	0.1412		0.8588	
^&crucialness>=0.0912698413&subjective_quality(No because of Odour, No because of Taste, No because of Colour)^&pay recoded(No)&l...	0.1992		0.8008	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)^&water_source_clean recoded(Borehole)&local_po...	0.2026		0.7974	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)&clean_adm1(Edo, Oyo, Benue, Anambra, Ekiti, Osu...	0.2103		0.7897	
^&crucialness>=0.0912698413&subjective_quality(No because of Odour, No because of Taste, No because of Colour)&clean_adm1(Akwa Ibo...	0.2368		0.7632	
clean_adm1(Akwa Ibom, Abia, Kogi, Ondo, Cross River, Bayelsa, Delta, Imo, Ogun, Kebbi, Ekiti, Edo, Anambra, Gombe, Benue, Sokoto, Plateau, ...	0.2595		0.7405	
^&crucialness>=0.0912698413&subjective_quality(No because of Odour, No because of Taste, No because of Colour)&clean_adm1(Akwa Ibo...	0.2610		0.7390	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)&clean_adm1(Abia, Kogi, Ondo, Gombe, Kebbi, Ogu...	0.2730		0.7270	
^&crucialness>=0.0912698413&subjective_quality(No because of Odour, No because of Taste, No because of Colour)&clean_adm1(Akwa Ibo...	0.2731		0.7269	
^&crucialness<0.0912698413&crucialness>=0.0200892857&local_population_1km recoded(1 - 1000)&is_urban(True)&clean_adm1(Abia, Keb...	0.2745		0.7255	
^&crucialness<0.0912698413&crucialness>=0.0200892857&local_population_1km recoded(1 - 1000)&is_urban(True)^&clean_adm1(Plateau,...	0.2767		0.7233	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)&clean_adm1(Abia, Kogi, Ondo, Gombe, Kebbi, Ogu...	0.2939		0.7061	
^&crucialness>=0.0374220374&subjective_quality(Acceptable quality)^&water_tech_clean recoded(Hand Pump)&served_population recode...	0.3105		0.6895	
^&crucialness>=0.0912698413&subjective_quality(No because of Odour, No because of Taste, No because of Colour)&clean_adm1(Akwa Ibo...	0.3123		0.6877	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)^&water_source_clean recoded(Borehole)&clean_ad...	0.3412		0.6588	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)&clean_adm1(Abia, Kogi, Ondo, Gombe, Kebbi, Ogu...	0.3521		0.6479	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)^&water_source_clean recoded(Borehole)&local_po...	0.3616		0.6384	
^&crucialness>=0.0912698413&subjective_quality(Acceptable quality)&pay recoded(No)^&water_source_clean recoded(Others)&water_tech...	0.3690		0.6310	
^&crucialness>=0.0912698413&subjective_quality(No because of Odour, No because of Taste, No because of Colour)&clean_adm1(Edo, Kwa...	0.3817		0.6183	

Figure 21 Decision Tree Leaf Report

4.6. Bootstrap Forest

A bootstrap forest was carried out with the results as follows. The resulting model is a decision tree with >600 splits. The model has a misclassification rate of 0.25, a non-functional prediction rate of 0.59, and a functional prediction rate of 0.84. The false positive error is 0.4 and false negative error is 0.15.

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2444	0.2005	0.1999	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3772	0.3182	0.3174	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5021	0.5313	0.5316	$\sum$ -Log(ρ[j])/n
RASE	0.4089	0.4237	0.4235	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.3573	0.3699	0.3691	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.2471	0.2739	0.2738	$\sum$ (ρ[j]≠ρMax)/n
N	32744	24558	24558	n

Confusion Matrix

Training

Actual STATUS recoded	Predicted Count Functional	Non-functional
Functional	17206	3066
Non-functional	5025	7447

Actual STATUS recoded	Predicted Rate Functional	Non-functional
Functional	0.849	0.151
Non-functional	0.403	0.597

Validation

Actual STATUS recoded	Predicted Count Functional	Non-functional
Functional	12594	2609
Non-functional	4117	5238

Actual STATUS recoded	Predicted Rate Functional	Non-functional
Functional	0.828	0.172
Non-functional	0.440	0.560

Test

Actual STATUS recoded	Predicted Count Functional	Non-functional
Functional	12671	2533
Non-functional	4192	5162

Actual STATUS recoded	Predicted Rate Functional	Non-functional
Functional	0.833	0.167
Non-functional	0.448	0.552

Figure 22 Bootstrap Forest Model Results

From the prediction profiler, water systems with low crucialness, high local population, require payment and other water sources have higher probability of non-functionality. Low crucialness may mean lower priority for maintenance or rehabilitation. High local population may translate to higher usage, and if facility is not designed for it, may lead to insufficiency. As for water sources, it may be because wells and springs may require higher maintenance than boreholes.

Figure 23 Bootstrap Model Prediction Profiler

4.7. Model Comparison

The table below compares the results from the different models generated. The decision tree and bootstrap forest models had lower misclassification rates between 0.25 – 0.30, higher prediction of true positives of 0.55 – 0.62 and higher area under ROC curves of 0.77 to 0.83. The decision tree, however, has higher Type I error of 0.23.

Table 1 Comparison of results from all models

			Positive: Non-Functional		Negative: Functional		AUC
Model		Misclassification Rate	True Positive	False Positive	True Negative	False Negative	
Nominal Logistic Regression	TR	0.31	0.48	0.17	0.83	0.52	0.74
	VA	0.30	0.47	0.17	0.83	0.53	0.74
	TE	0.30	0.47	0.17	0.83	0.53	0.74
Stepwise (Forward) Log. Reg.	TR	0.31	0.47	0.17	0.83	0.53	0.74
	VA	0.31	0.47	0.17	0.83	0.53	0.74
	TE	0.31	0.47	0.17	0.83	0.53	0.74
Stepwise (Backward) Log. Reg.	TR	0.31	0.47	0.17	0.83	0.53	0.74
	VA	0.31	0.47	0.17	0.83	0.53	0.74
	TE	0.31	0.47	0.17	0.83	0.53	0.74
Decision Tree	TR	0.29	0.62	0.23	0.77	0.38	0.77
	VA	0.30	0.61	0.23	0.77	0.39	0.76
	TE	0.29	0.60	0.23	0.77	0.40	0.76
Bootstrap Forest	TR	0.25	0.60	0.15	0.85	0.40	0.83
	VA	0.27	0.56	0.17	0.83	0.44	0.79
	TE	0.27	0.55	0.17	0.83	0.45	0.79

In the model comparison, the profit matrix was specified by updating the probability threshold for non-functional levels to 0.38, estimated from the percentage of non-functionality in the dataset.

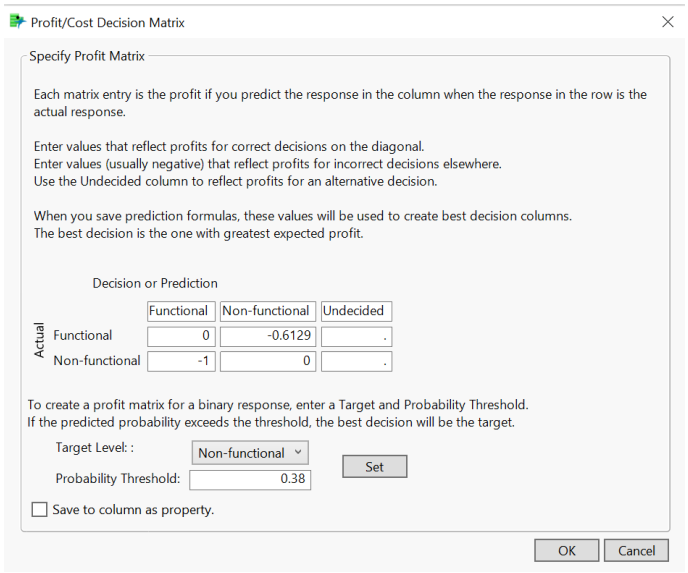


Figure 24 Probability Threshold Setting

After updating the profit matrix, the prediction for non-functionality improves for all the models. However, the Type I and Type II errors also increases. In general, based on the performance metrics, the decision tree and bootstrap forest models have higher sensitivity and accuracy than the logistic regression models. Some of the specificity and precision are on par.

		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

	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
Validation	**Method**	**TP**	**FN**	**TN**	**FP**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**	**F1**	**MCC**	**Profit**
	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
Testing	**Method**	**TP**	**FN**	**TN**	**FP**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**	**F1**	**MCC**	**Profit**
	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

Figure 25 Results after modifying probability threshold

Between the decision tree and bootstrap forest, the latter seems to outperform the decision tree in all aspects. 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 a sign of overfitting in the bootstrap forest model.

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.

	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

Comparing complexity, the decision tree model is less complex with 101 splits, compared to the bootstrap forest with over 600 splits. Hence, balancing performance metrics, complexity, and making sense of variables, the decision tree model seems to be the most suitable out of the 5 models.

5. Conclusion

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.

6. Recommendation

Based on the analysis of the models, the following recommendations are generated:

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.

b. 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.

c. 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.

d. 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.

7. Appendix

7.1. Appendix A: Fields in Original Dataset

Name	Label	Role	Level
row_id	Row number	ID	Continuous
source	Name of organization collecting the data record	Text	Nominal
lat_deg	Latitude of Water system	Input	Continuous
lon_deg	Longitude of Water system	Input	Continuous
report_date	Date that data was collected	Input	Continuous
status_id	Identify if any water is available on the day of visit	Text	Nominal
water_source_clean	Describe the water source	Text	Nominal
water_source_category	Describe the water source	Text	Nominal
water_tech_clean	Describe the system being used to transport the water	Text	Nominal
water_tech_category	Describe the system being used to transport the water	Text	Nominal
facility_type	Identify if the system is improved	Text	Nominal
clean_country_name	Name of country	Text	Nominal
clean_adm1	Name of primary administrative division	Text	Nominal
clean_adm2	Name of secondary administrative division	Text	Nominal
clean_adm3	Name of tertiary administrative division	Text	Nominal
install_year	4-digit installation year	Input	Continuous
installer	Name of entity that installed the system	Text	Nominal
rehab_year	4-digit of most recent major rehabilitation	Input	Nominal
rehabilitator	Name of entity that completed the rehabilitation	Text	Nominal
management_clean	Name of entity that manages the water point	Text	Nominal
pay	Payment amount and basis	Text	Nominal
status	Physical condition of water point	Text	Nominal
fecal_coliform_presence	Results of e.coli or water quality test	Input	Nominal
fecal_coliform_value	Results of e.coli or water quality test	Input	Nominal
subjective_quality	Information of the water including taste, appearance, odour	Input	Nominal
activity_id	Unique ID for the specific water point infrastructure	ID	Nominal
scheme_id	Identifier for a small, piped scheme that connects multiple water points	ID	Nominal
wpdx_id	Identifier	ID	Nominal
notes	Additional information	Text	Nominal
orig_lnk	Link to the data record for a specific water point or full data set	URL	Nominal
photo_lnk	URL of a photograph of the water system	URL	Nominal
country_id	Identifier for country name	ID	Nominal
data_lnk	Link to dataset	URL	Nominal
public_data_source	Link to dataset	URL	Nominal
distance_to_primary_road	Distance to primary road	Input	Continuous
distance_to_secondary_road	Distance to secondary road	Input	Continuous
distance_to_tertiary_road	Distance to tertiary road	Input	Continuous
distance_to_city	Distance to city	Input	Continuous
distance_to_town	Distance to town	Input	Continuous
rehab_priority	Priority of rehabilitation	Input	Continuous
served_population	Population served by water system	Input	Continuous
local_population_1km	Local population within 1km of water system	Input	Continuous

crucialness	Crucialness of water point (i.e. are there alternative water points nearby?)	Input	Continuous
pressure	Pressure measurement of water system (i.e. it the waterpoint over or under-utilised?)	Input	Continuous
usage_capacity	Recommended maximum users per water point	Input	Continuous
is_urban	Is in an urban area defined by EU Human Settlement Database	Text	Nominal
latest_record	Whether data contains latest information	Text	Nominal
location_id	Location ID of water system	ID	Continuous
cluster_size	Number of times location appears on dataset	Input	Continuous
lat_deg_original	Latitude of water system	Input	Nominal
lon_deg_original	Longitude of water system	Input	Nominal
water_source	Describe the water source	Text	Nominal
water_tech	Describe the system being used to transport the water	Text	Nominal
country_name	Name of country	Text	Nominal
adm1	Name of primary administrative division	Text	Nominal
adm2	Name of secondary administrative division	Text	Nominal
adm3	Name of tertiary administrative division	Text	Nominal
clean_country_id	Country ID	ID	Nominal
management	Entity managing water system	Text	Nominal
created_timestamp	Data creation timestamp	Input	Continuous
updated_timestamp	Data update timestamp	Input	Continuous
New Georeferenced Column	Georeferenced Latitude and Longitude combined	Text	Nominal
lat_lon_deg	Latitude and Longitude combined	Text	Nominal
count	Number of records of this type	Input	Continuous
converted	List of columns by WPDx from original data source	Text	Nominal

7.2. Appendix B: Data Preparation Change Log

Item	Variable Name	Issue	Action
1	row_id	Data characterised as continuous	Change to nominal
2	usage_capacity	Data characterised as continuous	Change to nominal
3	latest_record	Some data not updated, may not be representative of actual	Hide and exclude 752 rows with 'latest_record' = False
4	row_id	Row number irrelevant to analysis	Hide and exclude variable
5	source	Data source irrelevant to analysis	Hide and exclude variable
6	report_Date	Data source irrelevant to analysis	Hide and exclude variable
7	facility_type	Data source irrelevant to analysis	Hide and exclude variable
8	clean_country_name	Data source irrelevant to analysis	Hide and exclude variable
9	activity_id	Water point ID irrelevant to analysis	Hide and exclude variable
10	scheme_id	Water connection ID irrelevant to analysis	Hide and exclude variable
11	wpdx_id	WPDx ID irrelevant to analysis	Hide and exclude variable
12	notes	Additional information irrelevant to analysis	Hide and exclude variable
13	orig_lnk	URL link irrelevant to analysis	Hide and exclude variable
14	photo_lnk	URL link irrelevant to analysis	Hide and exclude variable
15	country_id	Country ID irrelevant to analysis	Hide and exclude variable
16	data_lnk	URL link irrelevant to analysis	Hide and exclude variable
17	public_data_source	URL link irrelevant to analysis	Hide and exclude variable
18	latest_record	Record latestness irrelevant to analysis	Hide and exclude variable
19	location_id	Duplicate information, able to use latitude and longitude data	Hide and exclude variable
20	cluster_size	Number of times record appears irrelevant to analysis	Hide and exclude variable
21	lat_deg_original	Duplicate information	Hide and exclude variable
22	lon_deg_original	Duplicate information	Hide and exclude variable
23	water_source	Duplicate information	Hide and exclude variable
24	water_tech	Duplicate information	Hide and exclude variable
25	country_name	Country name irrelevant to analysis	Hide and exclude variable
26	adm1	Duplicate information, clean data available	Hide and exclude variable
27	adm2	Duplicate information, clean data available	Hide and exclude variable
28	adm3	Duplicate information, clean data available	Hide and exclude variable
29	clean_country_id	Country ID irrelevant to analysis	Hide and exclude variable
30	management	Management entity irrelevant to analysis	Hide and exclude variable
31	created_timestamp	Data creation time irrelevant to analysis	Hide and exclude variable
32	updated_timestamp	Data updated time irrelevant to analysis	Hide and exclude variable
33	New Georeferenced Column	Duplicate information, contains latitude and longitude data	Hide and exclude variable
34	lat_lon_deg	Duplicate information, contains latitude and longitude data	Hide and exclude variable
35	count	Number of records of this type irrelevant to analysis	Hide and exclude variable
36	converted	List of columns converted irrelevant to analysis	Hide and exclude variable
37	install_year	More than 30% of data (39784 rows) is missing	Hide and exclude variable

38	rehab_priority	More than 30% of data (53591 rows) is missing	Hide and exclude variable
39	clean_adm2	Contains too many categories (746)	Hide and exclude variable, use clean_adm1 instead
40	clean_adm3	Data is missing	Hide and exclude variable
41	installer	Data is missing	Hide and exclude variable
42	rehab_year	Data is missing	Hide and exclude variable
43	rehabilitator	Data is missing	Hide and exclude variable
44	management_clean	More than 30% of data (27352 rows) is missing	Hide and exclude variable
45	fecal_coliform_presence	Data is missing	Hide and exclude variable
46	fecal_coliform_value	Data is missing	Hide and exclude variable
47	pay	Contains too many categories (42)	Recode into 'pay recoded' - 2 categories (Yes and No)
48	pay	Duplicate with pay recoded	Hide and exclude variable
49	status	Contains too many categories (881)	Recode into 'STATUS recoded' 2 categories (Functional and Non-Functional)
50	status	Duplicate with STATUS recoded	Hide and exclude variable
51	served_population	Contains 274 rows with missing data, may affect model analysis	Hide and exclude rows with missing data
52	local_population_1km	Contains 274 rows with missing data, may affect model analysis	Hide and exclude rows with missing data
53	crucialness	Contains 5616 rows with missing data, may affect model analysis	Hide and exclude rows with missing data
54	pressure	Contains 6443 rows with missing data, may affect model analysis	Hide and exclude rows with missing data
55	water_tech_clean	Contains 1224 rows with missing data, may affect model analysis	Hide and exclude rows with missing data
56	water_tech_category	Contains 1224 rows with missing data, may affect model analysis	Hide and exclude rows with missing data
57	usage_capacity	Contains 892 rows with missing data, may affect model analysis	Hide and exclude rows with missing data
58	served_population	Distribution is uneven with > 99% water points serving < 20000 population	Recode the columns into 'served_population_recoded' 5 categories – '0', '1-1000, 1001-10000, 10001 – 100000 and > 100000)
59	served_population	Duplicate with recoded column	Hide and exclude variable
60	local_population_1km	Distribution is uneven with > 97.5% water points serving < 20000 population	Recode the columns into 'local_population_1km_recoded' 5 categories – '0', '1-1000, 1001-10000, 10001 – 100000 and > 100000)
61	local_population_1km	Duplicate with recoded column	Hide and exclude variable
62	pressure	Correlated with served_population, may serve as redundancy	Hide and exclude variable
63	status_id	Binary predictor unstable in model	Hide and exclude variable
65	water_source_clean	Biasness in model may be due to uneven category distribution	Recode 2 categories – 'Borehole' and 'Others'
65	water_source_clean	Duplicate with water_source_clean	Hide and exclude variable
66	water_source_category	Similar to water_source_clean data, resulting in zeroed error	Hide and exclude variable
67	water_tech_clean	Biasness in model may be due to uneven category distribution	Recode 2 categories – 'Handpump' and 'Others'
68	water_tech_clean	Duplicate with water_tech_clean_recoded	Hide and exclude variable

69	water_tech_category	Similar to water_tech_clean data, resulting in zeroed error	Hide and exclude variable