ASSIGNMENT 3

ISSS602 – Data Analytics Lab

DATA ANALYTICS FOR GOOD
Prediction of Nonfunctional Water Points in Nigeria
10-Nov-2021

Le Vu Anh Phuong (Joshua)
vaple.2021@mitb.smu.edu.sg

Contents

1. OVERVIEW

According to UN-Water, 1.8 billion people will face serious water scarcity by 2025 with developing countries bearing the greatest risk. Over the past 40 years, there have been many initiatives to provide the much-needed resources in the form of building and running of thousands of water points in rural communities, especially in sub-Saharan African nations. However, there are still ongoing issues of waterpoints being nonfunctional with various reasons from dry/low-yielding to being marked with tastes, appearance and odour problems (Liddle, 2017).

2. OBJECTIVES

The objective of this study is to build a prediction model for the nonfunctional waterpoints for the country of Nigeria. The model aims at targeting the right attention and limited resources to where they are needed the most.

3. DATA

3.1 Data Used

The data source to build this model is from the global initiative called Water Point Data Exchange. This project is set out to collect waterpoints data from rural areas and share it via a cloud-based data library in a standardized manner for efficient analysis.

3.2 Data Preparation

3.2.1 Defining the Target Variable

The "status" column provides some details whether the water points are functional or nonfunctional, with associated reasons (technical and non-technical reasons). To achieve the level of focused required by the stakeholder, recoding is performed only with the following Status descriptions. The rest are excluded due to being overly detailed and account for a very small amount of data points. The resultant response variable column is named **"status_recode".**

- Functional (and in use): Recoded to be "Functional"
- Non-functional technical breakdown + Functional (but not in use) Technical Breakdown: Recoded to be "Nonfunctional"
- Non-functional Dry/low-yielding: Recoded to be "Nonfunctional "

There are 79,414 rows remaining, upon 89,447 rows in total (~90%)

Frequencies

Level	Count	Prob
Total	55202	1.00000
Functional (and in use)	48317	0.87528
Functional and in use but in bad shape	1718	0.03112
Functional (but not in use) Technical breakdown	1513	0.02741
Functional (but not in use)	1495	0.02708
Functional (but not in use) Under rehabilitation	559	0.01013
Functional and in use but in bad shape Technical breakdown	388	0.00703
Functional (but not in use) Dry/low-yielding	276	0.00500
Functional and in use but in bad shape Dry/low-yielding	231	0.00418
Functional (but not in use) Water quality	153	0.00277
Functional (but not in use) New under construction	99	0.00179
Functional and in use but in bad shape Water quality	67	0.00121
Functional and in use but in bad shape Under rehabilitation	50	0.00091
Functional (but not in use) Closer alternative (improved) source	27	0.00049
Functional and in use but in bad shape N/A, currently Functional (and in use)	23	0.00042
Functional (but not in use) Cheaper alternative (improved) source	16	0.00029
Functional (but not in use) Free (unimproved) source	14	0.00025
Functional (but not in use) N/A, currently Functional (and in use)	14	0.00025
Functional and in use but in bad shape New under construction	14	0.00025
Functional (but not in use) Silted (dams/pans only)	7	0.00013
Functional (but not in use) Thief Stolen	7	0.00013
Functional and in use but in bad shape Closer alternative (improved) source	6	0.00011
Functional and in use but in bad shape Free (unimproved) source	5	0.00009
Functional (but not in use) Lack of power supply	4	0.00007
Functional (but not in use) No power supply	4	0.00007

N Missing 74
200 Levels

Frequencies

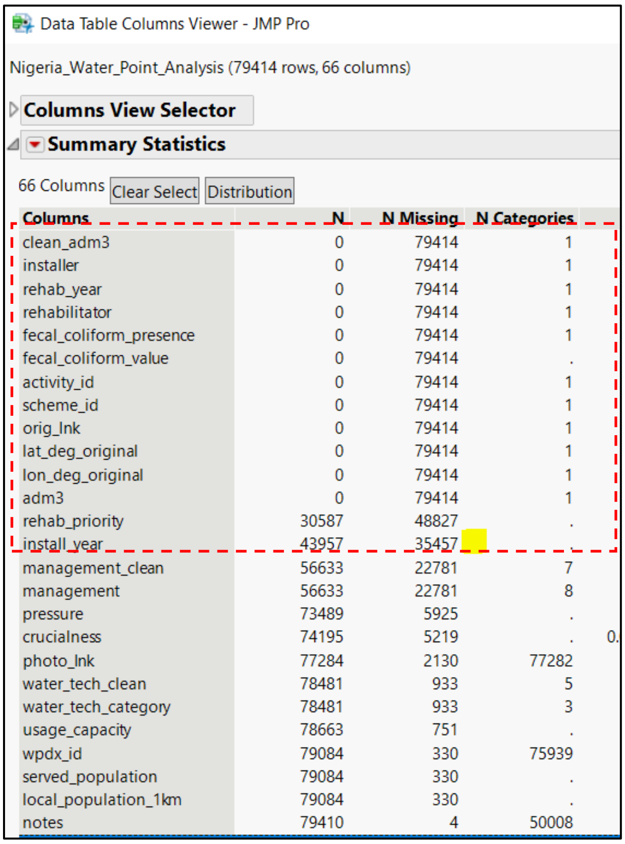
Level	Count	Prob
Total	34167	1.00000
Non-functional Technical breakdown	27055	0.79185
Non-functional Dry/low-yielding	2529	0.07402
Non-functional Under rehabilitation	1404	0.04109
Non-functional New under construction	1154	0.03378
Non-functional Water quality	347	0.01016
Non-functional Cheaper alternative (improved) source	112	0.00328
Non-functional Closer alternative (improved) source	107	0.00313
Non-functional Silted (dams/pans only)	94	0.00275
Non-functional Free (unimproved) source	68	0.00199
Non-functional FAULTY PUMP	58	0.00170
Non-functional abandoned	54	0.00158
Non-functional N/A, currently Functional (and in use)	51	0.00149
Non-functional Abandoned	39	0.00114
Non-functional Abandoned Project	24	0.00070
Non-functional Abortive	24	0.00070
Non-functional uncompleted project	22	0.00064
Non-functional Uncompleted Project	18	0.00053
Non-functional Abandoned project	16	0.00047
Non-functional ABANDON	15	0.00044
Non-functional Vandalised	15	0.00044
Non-functional ABANDONED	14	0.00041
Non-functional Totally Damaged	14	0.00041
Non-functional uncompleted	14	0.00041
Non-functional	10	0.00029

N Missing 0
692 Levels

3.2.2 Data Wrangling

Handling Missing Data

The following columns are excluded due to missing data of >30% of total data **Figure 1**. Note that "install_year" is related to the age of the water points and hence may be of importance to predict their conditions. However, in this study, due to the large amount of missing data, this field is excluded and hence may contribute to the model limitations.


Data Table Columns Viewer - JMP Pro

Nigeria_Water_Point_Analysis (79414 rows, 66 columns)

Columns View Selector

Summary Statistics

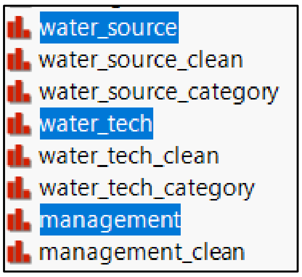
66 Columns Clear Select Distribution

Columns	N	N Missing	N Categories
clean_adm3	0	79414	1
installer	0	79414	1
rehab_year	0	79414	1
rehabilitator	0	79414	1
fecal_coliform_presence	0	79414	1
fecal_coliform_value	0	79414	.
activity_id	0	79414	1
scheme_id	0	79414	1
orig_lnk	0	79414	1
lat_deg_original	0	79414	1
lon_deg_original	0	79414	1
adm3	0	79414	1
rehab_priority	30587	48827	.
install_year	43957	35457	.
management_clean	56633	22781	7
management	56633	22781	8
pressure	73489	5925	.
crucialness	74195	5219	.
photo_lnk	77284	2130	77282
water_tech_clean	78481	933	5
water_tech_category	78481	933	3
usage_capacity	78663	751	.
wpdx_id	79084	330	75939
served_population	79084	330	.
local_population_1km	79084	330	.
notes	79410	4	50008

Figure 1: Handling Missing Data

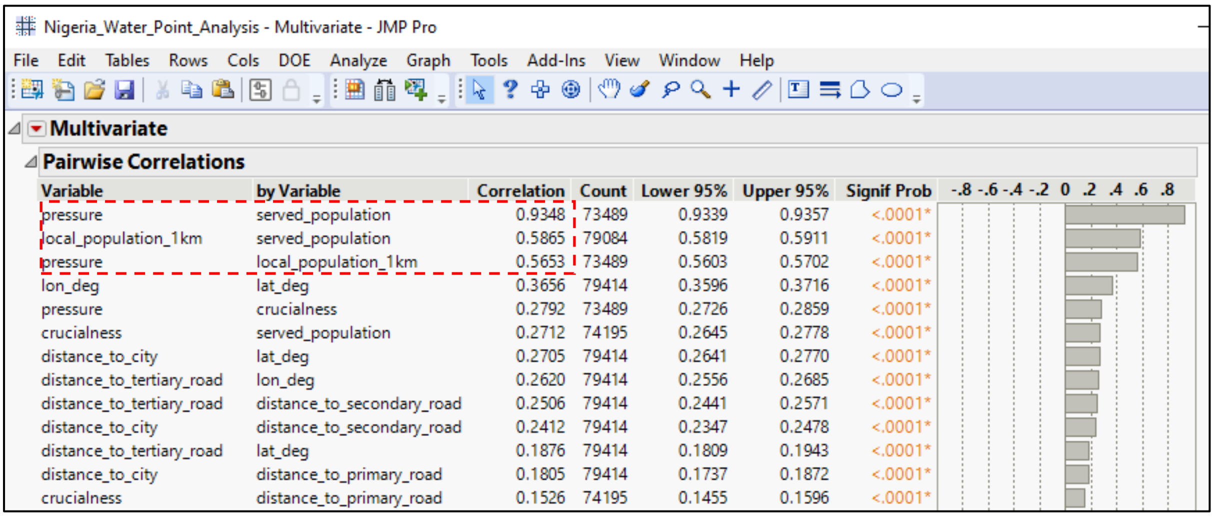
Handling Old Data Fields

As describe from the Data Standard document of the data provider, fields with "_clean" suffix are the result of data cleaning exercise done previously to produce this dataset. Hence, they are retained and the original fields (without this suffix) are excluded.



Handling Correlated Data

Next, multivariate analysis is performed to identify any issue of multi-collinearity of continuous variables. It can be seen in **Figure 2** that pressure and local_population_1km are strongly and moderately correlated to served_population. As served_population is a demographic data of the communities immediately helped by the water points, hence it is kept in this study while the other 2 are excluded.

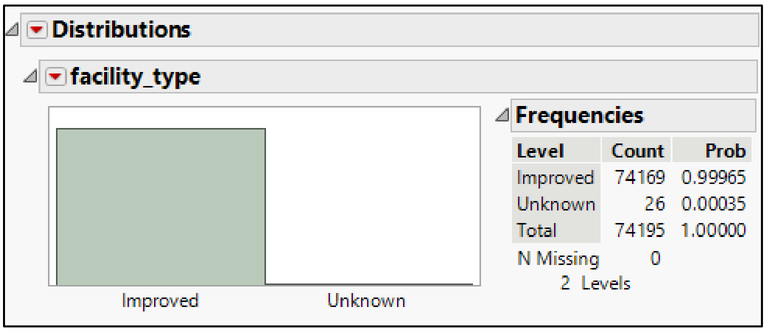


Variable	by Variable	Correlation	Count	Lower 95%	Upper 95%	Signif Prob
pressure	served_population	0.9348	73489	0.9339	0.9357	<.0001*
local_population_1km	served_population	0.5865	79084	0.5819	0.5911	<.0001*
pressure	local_population_1km	0.5653	73489	0.5603	0.5702	<.0001*
lon_deg	lat_deg	0.3656	79414	0.3596	0.3716	<.0001*
pressure	crucialness	0.2792	73489	0.2726	0.2859	<.0001*
crucialness	served_population	0.2712	74195	0.2645	0.2778	<.0001*
distance_to_city	lat_deg	0.2705	79414	0.2641	0.2770	<.0001*
distance_to_tertiary_road	lon_deg	0.2620	79414	0.2556	0.2685	<.0001*
distance_to_tertiary_road	distance_to_secondary_road	0.2506	79414	0.2441	0.2571	<.0001*
distance_to_city	distance_to_secondary_road	0.2412	79414	0.2347	0.2478	<.0001*
distance_to_tertiary_road	lat_deg	0.1876	79414	0.1809	0.1943	<.0001*
distance_to_city	distance_to_primary_road	0.1805	79414	0.1737	0.1872	<.0001*
crucialness	distance_to_primary_road	0.1526	74195	0.1455	0.1596	<.0001*

Figure 2: Multivariate Analysis of Continuous Variables

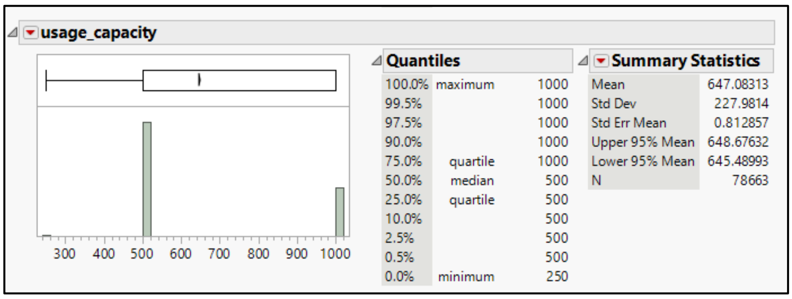
Handling highly Skewed Data and Overly Detailed Categorical Data

- Vast majority of the water points are under "Improved" facility type, hence this variable has no predicting property and can be excluded.



Level	Count	Prob
Improved	74169	0.99965
Unknown	26	0.00035
Total	74195	1.00000

- Usage_capacity is essentially categorical data with 3 levels (250, 500, 1000), hence it is changed from continuous to normal data type.



100.0%	maximum	1000
99.5%		1000
97.5%		1000
90.0%		1000
75.0%	quartile	1000
50.0%	median	500
25.0%	quartile	500
10.0%		500
2.5%		500
0.5%		500
0.0%	minimum	250

Mean	647.08313
Std Dev	227.9814
Std Err Mean	0.812857
Upper 95% Mean	648.67632
Lower 95% Mean	645.48993
N	78663

- The "pay" column indicates the payment scheme of the water points. There are a lot of schemes throughout the country but essentially, they belong to fee-paying or free categories.

Frequencies

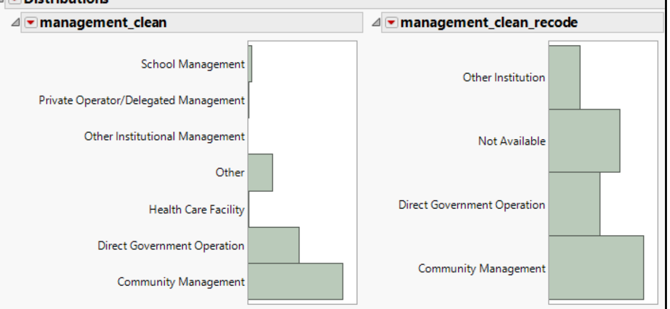
Level	Count	Prob
Total	79414	1.00000
No	73366	0.92384
Yes Point of collection	2884	0.03632
Yes At breakdown	2597	0.03270
Yes Periodic Levy	525	0.00661
Yes For Maintenance	2	0.00003
Yes household	2	0.00003
Yes monthly	2	0.00003
Yes Monthly	2	0.00003
Yes Pry Sch	2	0.00003
Yes weekly	2	0.00003
Yes #10/20litre	1	0.00001
Yes by caretakers	1	0.00001
Yes Commatee	1	0.00001
Yes COMMIMITY	1	0.00001
Yes DAILY	1	0.00001
Yes daily charges	1	0.00001
Yes During Power outage	1	0.00001
Yes Everyday	1	0.00001
Yes fifty naira monthly for maintenance	1	0.00001
Yes For fuel	1	0.00001
Yes For Fuel	1	0.00001
Yes FOR MAINTANCE	1	0.00001
Yes for maintenance	1	0.00001
Yes FOR MANTANEANCE AND FUEL	1	0.00001
Yes for repair	1	0.00001
Yes Instantly	1	0.00001
Yes LGA	1	0.00001
Yes MAINTENANCE AND FUEL	1	0.00001
Yes National Union Of Road Transport Workers repairs it	1	0.00001
Yes people outside the community pay during dry season	1	0.00001
Yes POINT OF COLLECTION FOR COMMERCIAL TRUCKS BUT FREE FOR COMMUNITY	1	0.00001

N Missing 0

Hence all the "Yes" categories are recoded into a single "Yes" category, while the "No" category is kept as original. The resultant column is named "**pay_recoded**"

- It is observed in the management_clean field that apart from "Community Management" and "Direct Government Operation", the other types of management schemes for the water points are in small numbers. Hence, they are grouped into a single category of "Other Institution" via recoding. In addition, blanks are categorised as "Not Applicable" meaning there is no management scheme for these water points. The resultant field is named "management_clean_recoded".

Distributions



Frequencies (management_clean)

Level	Count	Prob
Community Management	30235	0.53388
Direct Government Operation	16376	0.28916
Health Care Facility	407	0.00719
Other	7967	0.14068
Other Institutional Management	124	0.00219
Private Operator/Delegated Management	195	0.00344
School Management	1329	0.02347
Total	56633	1.00000

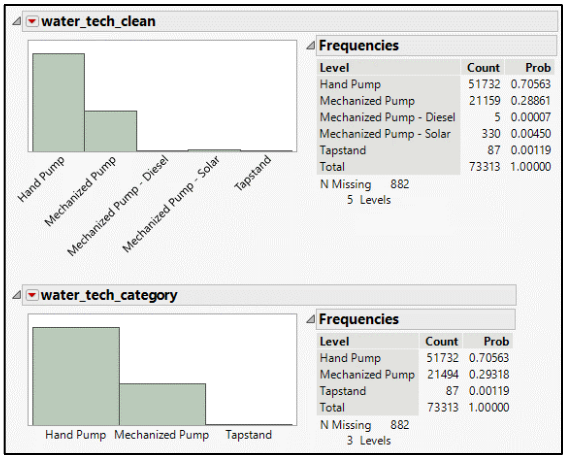
N Missing 22781
7 Levels

Frequencies (management_clean_recode)

Level	Count	Prob
Community Management	30235	0.38073
Direct Government Operation	16376	0.20621
Not Available	22781	0.28686
Other Institution	10022	0.12620
Total	79414	1.00000

N Missing 0
4 Levels

- It can be observed that water_tech_clean only provides more elaborated descriptions for the "Mechanized Pump" category under the field "water_tech_category". In addition, "Tapstand" is a form of manual pump. Hence, it is grouped with "Hand Pump". Thus, this study will only use water_tech_category with the recoding done to group the tapstand above. The resultant field is named "water_tech_category_recode"



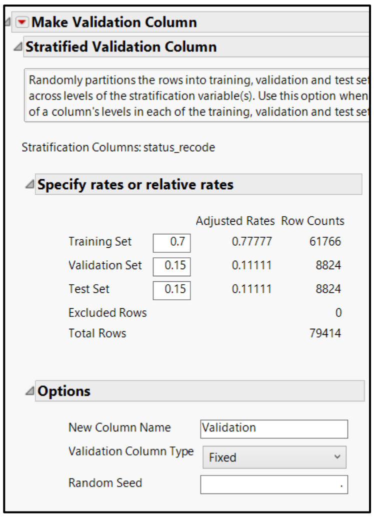
- The distance columns are reformatted as below for easier reading of values.

distance_to_primary_road	distance_to_secondary_road	distance_to_tertiary_road	distance_to_city	distance_to_town
1,675	16,461	4,746	81,958	31,976
3,610	14,530	2,381	83,592	37,873

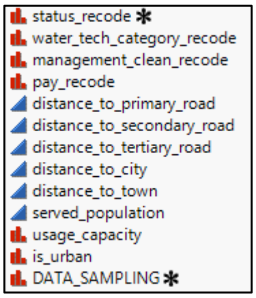
<u>Exclusion of Other Fields</u>

- "crucialness" and "subjective_quality": there are no explanation of these fields in the data source. Based on the title, they may be related to the status of the water points (response variable) yet not flagged by multivariate test as the relationship may not be linear.
- Geographical and administration fields: They may be useful in real life, however, as the understanding of differences between these locations are not available in this study, they are excluded.
  - "lat_deg" and "lon_deg": these fields refer to the latitude and longitude positions of the water points.
  - "amd1" (37 levels of main provinces), "amd2" (746 levels of smaller towns and villages).
- Other fields that contain data administration information such as "timestamps", "data_lnk", "photo_lnk" are excluded as they should not have effects to the response variable.

The final step of data preparation is to create validation column named DATA_SAMPLING based on status_recode for predictive modelling purpose. More data is allocated to the training set while the remaining 2 sets still have more than 500 rows per predictors (~9000 rows).



<u>The final fields are listed below with 11 predictors:</u>



4. ANALYSIS

4.1 Logistic regression modelling

As the response variable is categorical in nature, logistic regression is first applied. Using Logistic Regression platform on all remaining variables with status_recode as the response variable (Target = Nonfunctional) (**Figure 3**), we obtain the following result summarized in **Figure 4**.

Firstly, the model fit is examined. The **Whole model test** is testing $H0$ of “The logistic model is NOT useful to explain the data” and “Ha: The logistic model is useful to explain the data”. From the p-value <0.0001, H0 is rejected, the model is significant in explaining the response variable. In addition, the **Lack of Fit test** is testing $H0$ of “The model is adequate to explain the data” and Ha of “The model is inadequate.

From the p-value of <0.0001, H0 is rejected, the lack of fit Chi-square is significant, there is some significant benefit in introducing additional variables.

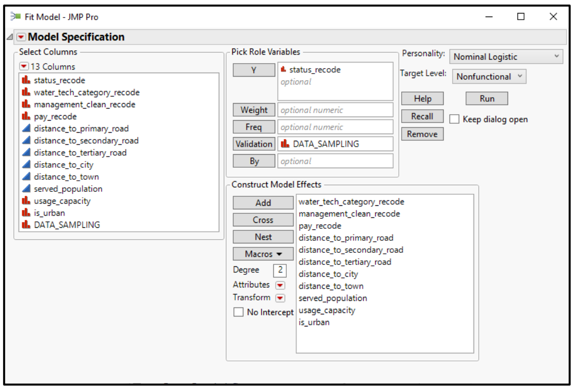
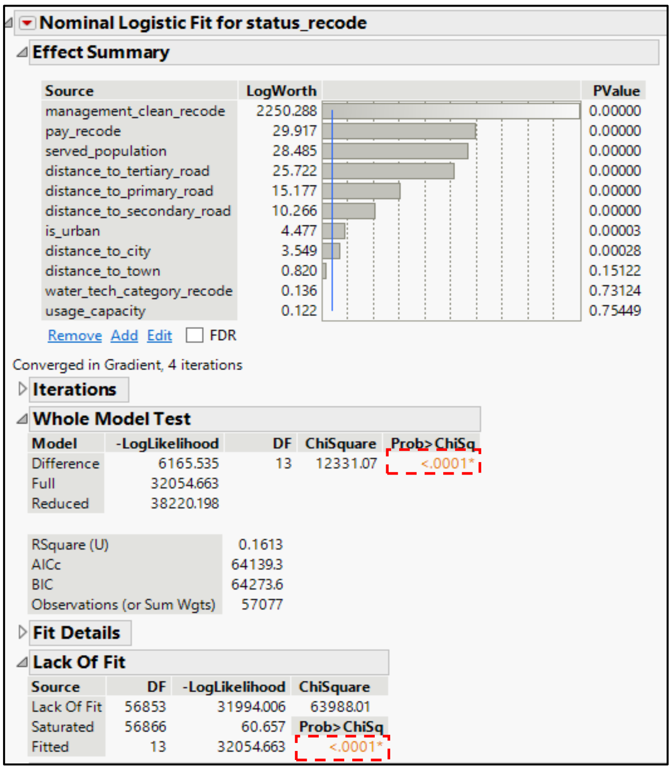


Figure 4: Logistic Regression Platform



Nominal Logistic Fit for status_recode

Effect Summary

Source	LogWorth	PValue
management_clean_recode	2250.288	0.00000
pay_recode	29.917	0.00000
served_population	28.485	0.00000
distance_to_tertiary_road	25.722	0.00000
distance_to_primary_road	15.177	0.00000
distance_to_secondary_road	10.266	0.00000
is_urban	4.477	0.00003
distance_to_city	3.549	0.00028
distance_to_town	0.820	0.15122
water_tech_category_recode	0.136	0.73124
usage_capacity	0.122	0.75449

Remove Add Edit FDR

Converged in Gradient, 4 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	6165.535	13	12331.07	<.0001*
Full	32054.663			
Reduced	38220.198			

RSquare (U)	0.1613
AICc	64139.3
BIC	64273.6
Observations (or Sum Wgts)	57077

Fit Details

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	56853	31994.006	63988.01
Saturated	56866	60.657	Prob>ChiSq
Fitted	13	32054.663	<.0001*

Figure 3: Logistic Regression Model Fit

From Parameter Estimate, a few columns show Biased and Zeroed flags (**Figure 5**), indicating that there are linear dependencies among model terms. By conducting interactive data exploration of their distribution, it can be seen that all the water points of high usage capacity of 1000 are equipped with mechanized pumps, while the lower capacity is by hand pumps (**Figure 6**). This strong relationship is the potential reason of the issue above, but not identified in multivariate analysis as they are categorical in nature. Hence we proceed with removing usage_capacity for this and future analysis.

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Biased	-0.8958277	0.2570452	12.15	0.0005*
water_tech_category_recode[Hand Pump]	Biased	-0.0891036	0.2572623	0.12	0.7291
management_clean_recode[Community Management]		-0.385038	0.0154744	619.12	<.0001*
management_clean_recode[Direct Government Operation]		-0.5763151	0.0190076	919.31	<.0001*
management_clean_recode[Not Available]		1.56526457	0.0166709	8815.7	<.0001*
pay_recode[No]		0.22288839	0.0198956	125.50	<.0001*
distance_to_primary_road		7.12549e-6	8.8027e-7	65.52	<.0001*
distance_to_secondary_road		6.52988e-6	9.9333e-7	43.21	<.0001*
distance_to_tertiary_road		-0.0000187	1.7825e-6	110.07	<.0001*
distance_to_city		-1.0043e-6	2.7691e-7	13.15	0.0003*
distance_to_town		7.93137e-7	5.5235e-7	2.06	0.1510
served_population		4.22735e-5	4.0325e-6	109.90	<.0001*
usage_capacity[250]	Biased	-0.1594112	0.5134718	0.10	0.7562
usage_capacity[500]	Zeroed	0	0	.	.
is_urban[False]		0.05710073	0.0137882	17.15	<.0001*

For log odds of Nonfunctional/Functional

Figure 5: Parameter Estimates

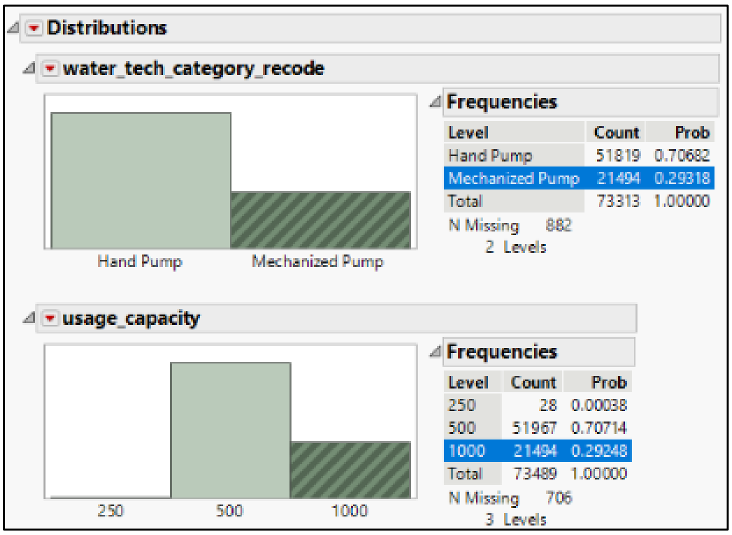


Figure 6: Interactive Data Exploration

After removing the above variable and rerunning the mode, the issue of bias and zeros are resolved as below:

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-0.8165127	0.0280241	848.91	<.0001*
water_tech_category_recode[Hand Pump]	-0.1689	0.0111407	229.85	<.0001*
management_clean_recode[Community Management]	-0.3849713	0.015473	619.03	<.0001*
management_clean_recode[Direct Government Operation]	-0.5762387	0.0190062	919.21	<.0001*
management_clean_recode[Not Available]	1.56524361	0.0166707	8815.7	<.0001*
pay_recode[No]	0.22300151	0.0198924	125.67	<.0001*
distance_to_primary_road	7.12733e-6	8.8026e-7	65.56	<.0001*
distance_to_secondary_road	6.53056e-6	9.9333e-7	43.22	<.0001*
distance_to_tertiary_road	-0.0000187	1.7825e-6	110.06	<.0001*
distance_to_city	-1.003e-6	2.7688e-7	13.12	0.0003*
distance_to_town	7.95169e-7	5.5232e-7	2.07	0.1500
served_population	4.22886e-5	4.0323e-6	109.98	<.0001*
is_urban[False]	0.057239	0.0137814	17.25	<.0001*

For log odds of Nonfunctional/Functional

Next, based on this revised model, the fit details are examined to assess the predictive performance of the model as it is applied on the training, validation and test data. Shown in **Figure 7**, the model shows decent fit with relatively small misclassification rate of ~0.26. from the Confusion Matrix, the model performs well predicting the true-negative (Functional) water points (~0.87) and not good at predicting the true-positive (Nonfunctional) water points (~0.53 - 0.55).

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1613	0.1551	0.1651	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2633	0.2542	0.2689	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5616	0.5658	0.5594	∑ -Log(ρ[j])/n
RASE	0.4332	0.4354	0.4320	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.3753	0.3778	0.3746	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2565	0.2601	0.2531	∑ (ρ[j]≠ρMax)/n
N	57077	8100	8136	n

Lack Of Fit

Parameter Estimates

Effect Likelihood Ratio Tests

Confusion Matrix

Training

Actual status_recode	Predicted Count Nonfunctional	Functional
Nonfunctional	12142	10231
Functional	4407	30297

Actual status_recode	Predicted Rate Nonfunctional	Functional
Nonfunctional	0.543	0.457
Functional	0.127	0.873

Validation

Actual status_recode	Predicted Count Nonfunctional	Functional
Nonfunctional	1691	1485
Functional	622	4302

Actual status_recode	Predicted Rate Nonfunctional	Functional
Nonfunctional	0.532	0.468
Functional	0.126	0.874

Test

Actual status_recode	Predicted Count Nonfunctional	Functional
Nonfunctional	1746	1451
Functional	608	4331

Actual status_recode	Predicted Rate Nonfunctional	Functional
Nonfunctional	0.546	0.454
Functional	0.123	0.877

Figure 7: Logistic Regression Model Predictive Performance

With the motivation to simplify the model as required by the stakeholder, step-wise method is conducted by the following setup (**Figure 8**)

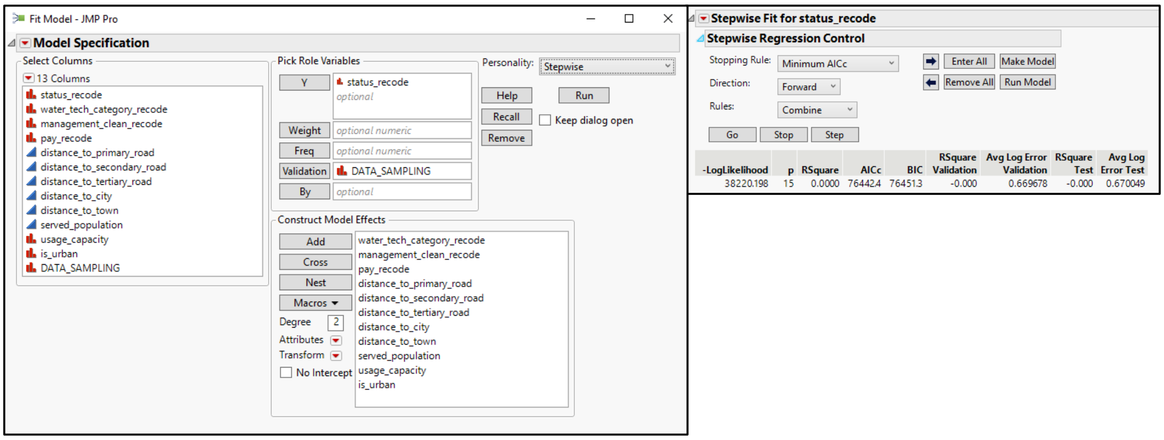


Figure 8: Step-wise Logistic Regression Setup

From the Step History (**Figure 9**), the lowest AIC and BIC are achieved at Step 11, corresponding to "distance_to_town" variable. This is in good agreement with the Parameter Estimates table in the previous logistic regression, where this variable is the only insignificant contributor with p-value of 0.15.

Step History

Step	Parameter	Action	L-R ChiSquare	"Sig Prob"	Entry ChiSquare	Entry "Sig Prob"	RSquare	p	AICc	BIC	RSquare Validation
1	management_clean_recode{Direct Government Operation&Other Institution&Community Management-Not Available}	Entered	11520.3	0.0000	11476	0	0.1507	2	64924.1	64942	0.1447
2	water_tech_category_recode{Hand Pump-Mechanized Pump}	Entered	242.3215	0.0000	244.646	3.8e-55	0.1539	3	64683.8	64710.6	0.1470
3	served_population	Entered	126.7212	0.0000	124.965	5.2e-29	0.1555	4	64559	64594.9	0.1494
4	pay_recode{Yes-No}	Entered	119.9027	0.0000	114.697	9.2e-27	0.1571	5	64441.1	64485.9	0.1519
5	management_clean_recode{Direct Government Operation&Other Institution-Community Management}	Entered	93.29639	0.0000	93.0241	5.2e-22	0.1583	6	64349.9	64403.6	0.1540
6	distance_to_primary_road	Entered	73.83871	0.0000	74.3529	6.5e-18	0.1593	7	64278	64340.7	0.1546
7	distance_to_tertiary_road	Entered	79.52091	0.0000	77.8811	1.1e-18	0.1603	8	64200.5	64272.1	0.1551
8	distance_to_secondary_road	Entered	44.04935	0.0000	44.3391	2.8e-11	0.1609	9	64158.4	64239	0.1554
9	is_urban{True-False}	Entered	15.74083	0.0001	15.6854	7.48e-5	0.1611	10	64144.7	64234.2	0.1555
10	distance_to_city	Entered	12.74908	0.0004	12.7262	0.00036	0.1613	11	64134	64232.4	0.1554
11	distance_to_town	Entered	1.933157	0.1644	1.9352	0.16419	0.1613	12	64134	64241.4	0.1551
12	management_clean_recode{Direct Government Operation-Other Institution}	Entered	0.593641	0.4410	0.59301	0.44126	0.1613	13	64135.4	64251.8	0.1551
13	Best	Specific	.	.	0.59301	0.44126	0.1613	11	64134	64232.4	0.1554

Figure 9: Step History of Step-Wise Method

By clicking "Make Model", JMP automatically removed the parameters #11 and #12. Running the logistic regression again based on the reduced predictor set we obtained the following report **Figure 10**. Examining the fit report and confusion matrix of the reduced model, it can be seen that the misclassification rate is slightly reduced. However, this is insignificant, and the confusion matrix prediction rates are almost unchanged from the unreduced regression model. Thus, step-wise method is not useful in this case to improve the model significantly.

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1613	0.1554	0.1651	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2632	0.2546	0.2689	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5616	0.5656	0.5594	∑ -Log(ρ[j])/n
RASE	0.4332	0.4354	0.4320	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3753	0.3778	0.3747	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2564	0.2601	0.2531	∑ (ρ[j]≠ρMax)/n
N	57077	8100	8136	n

Lack Of Fit

Parameter Estimates

Effect Likelihood Ratio Tests

Confusion Matrix

Training

Actual status_recode	Predicted Count Nonfunctional	Predicted Count Functional
Nonfunctional	12142	10231
Functional	4405	30299

Actual status_recode	Predicted Rate Nonfunctional	Predicted Rate Functional
Nonfunctional	0.543	0.457
Functional	0.127	0.873

Validation

Actual status_recode	Predicted Count Nonfunctional	Predicted Count Functional
Nonfunctional	1691	1485
Functional	622	4302

Actual status_recode	Predicted Rate Nonfunctional	Predicted Rate Functional
Nonfunctional	0.532	0.468
Functional	0.126	0.874

Test

Actual status_recode	Predicted Count Nonfunctional	Predicted Count Functional
Nonfunctional	1746	1451
Functional	608	4331

Actual status_recode	Predicted Rate Nonfunctional	Predicted Rate Functional
Nonfunctional	0.546	0.454
Functional	0.123	0.877

Figure 10: Fit Details of Reduced Regression Model

By removing the variable "distance_to_town" from the original regression model and rerun, we obtain the following Prediction Profiler to illustrate the direction of how different predictors affect the nonfunctionality of water points. This method of using Prediction Profiler is employed instead of saving the predictive formula as the later is mathematically inclined and not user-friendly to non-technical audience. From **Figure 11**, a few observations can be drawn:

- Mechanized pumps have higher nonfunctional proportion than hand pump. This could be due to their higher level of complexity for maintenance and repair, especially in rural areas.
- The absence of management scheme drastically increases the tendency of water points being nonfunctional. This is reasonable as local management structures lead to higher accountability and hence resources diverted to the maintenance and operation of the waterpoints.
- The fee-paying scheme (pay = “yes”) slightly improves the status of the water points with lower overall rate of nonfunctional.

- As distance to primary and secondary roads increases, the nonfunctional proportion increases, while the contrary is true for the distance to tertiary road. This could be because the higher distance to main roads (and consequently closer to lower tier roads), the more rural the communities. This may lead to higher scarcity of technical expertise and resources to operate and maintain the water points.
- The higher the served population, the more proportion of nonfunctional water points. This is reasonable because of more wear and tear are expected.
- Finally, whether the water points are located in urban area is not influencing their status significantly.

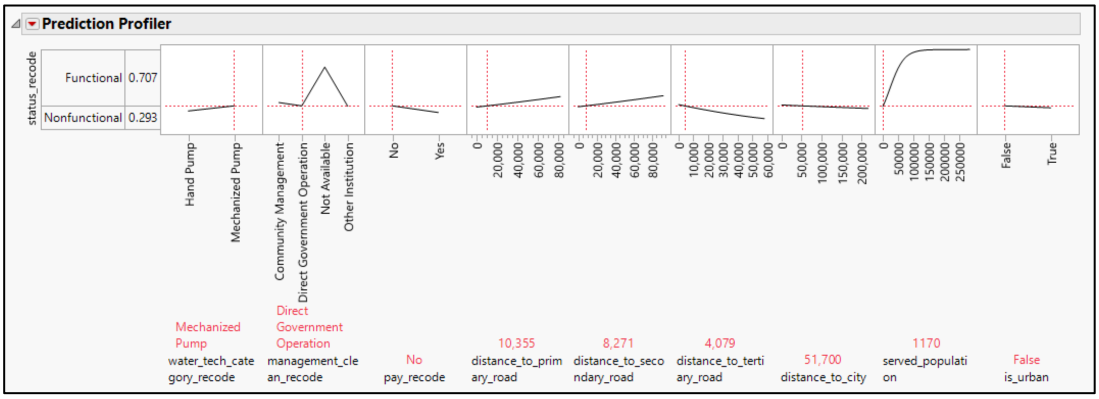
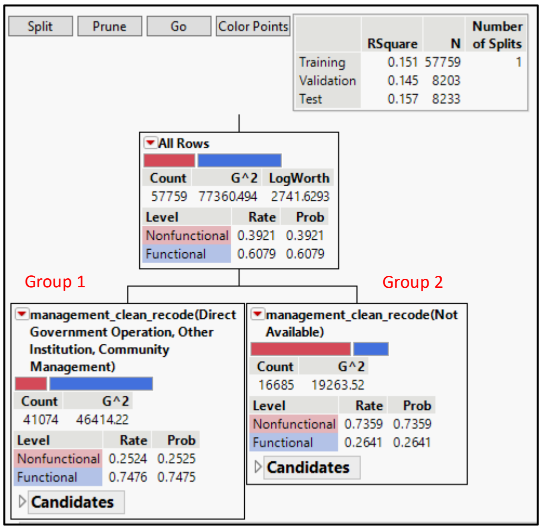


Figure 11: Logistic Regression Model Profiler

4.2 Decision Tree Modelling

As the above logistic regression analysis shows, there are some observations made on the variables that contribute to the nonfunctionality of the water points. However, the degree of application is limited as the stakeholders can only be provided with the univariable analysis of the profiler above or need to understand the mathematical model behind the Logit formula for meaningful discussions. Hence, the decision tree method is deployed to generate simpler rules to split the nonfunctional from functional water points in a more user-friendly manner.

Launching the Decision Tree platform, from the first split, we can clearly see that the presence of management structure is very significant in impacting the status of the water points.



By Pressing "Go" to let JMP grow and prune the tree and extend from the above 2 nodes, it can be seen that most nodes from the Group 1 above have Nonfunctional proportion of <40% whereas those on Group 2 have Nonfunctional proportion of approximately 65-80%.

As the tree has relatively many branches, the Leaf Report is used instead to identify important archetypes. We will focus the discussion on the first 6 groups with Nonfunctional proportion of >50% (Figure 12)

- Except from the first group, the remaining 5 have the same characteristic of no management structure.
- Among these 5 groups, mechanized pumps tend to have higher nonfunctional status
- The group with hand-pump that has higher nonfunctional proportion is serving a high population of >=1192 people without payment scheme.
- The first group with management structure but demonstrates exceptionally high nonfunctional rate, despite serving a low population and with payment scheme to service any defects is likely due to the distance to secondary road being too high. This is in contrast with another group with all factors being equal except this distance factor (Green Box) - at nonfunctional rate of ~0.2
- The only group that has no management structure but performs well with nonfunctional rate of 0.3272 is when it has short distance to primary road (Blue Box), compared with group #5 with higher distance and performs much more poorly at the rate of ~0.7
- It can be noted that distance to town and city are not contributing significantly to the splits in water point status.

Leaf Report

Response Prob

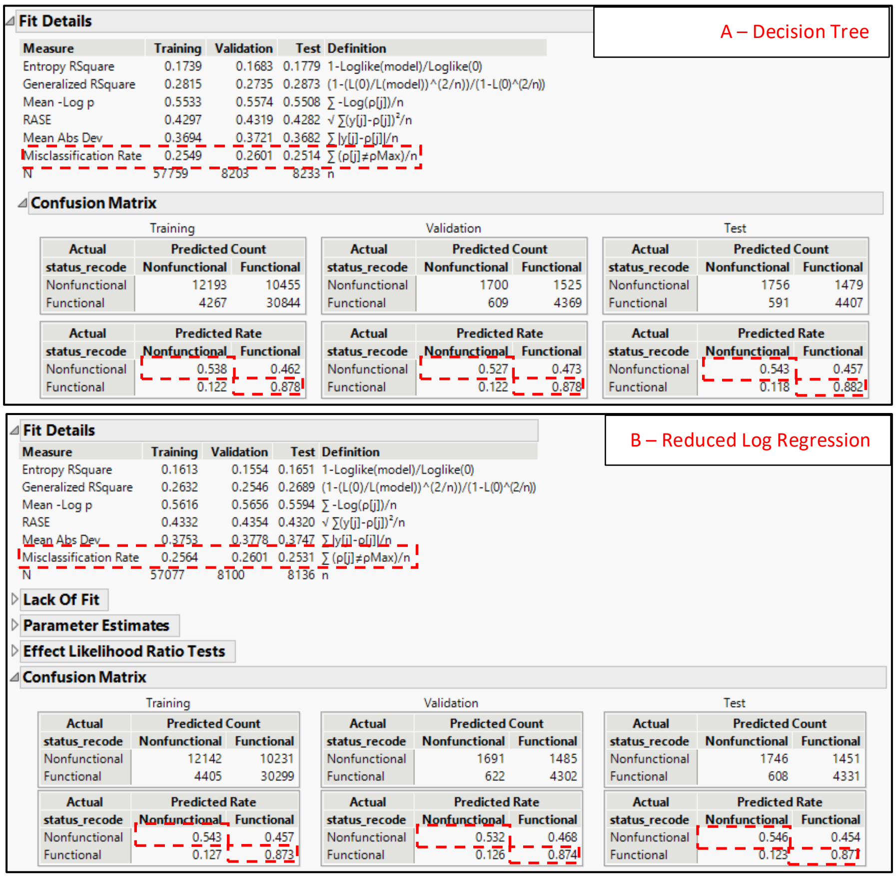
	Leaf Label	Nonfunctional	Functional
1	^&served_population<676&water_tech_category_recode(Hand Pump)&management_clean_recode(Direct Government Operation)&is_urban(False)&pay_recode(Yes)&distance_to_secondary_road>=22,183	0.8398	0.1602
2	management_clean_recode(Not Available)&water_tech_category_recode(, Mechanized Pump)&pay_recode(No)&is_urban(False)	0.8189	0.1811
3	management_clean_recode(Not Available)&water_tech_category_recode(Hand Pump)&served_population>=1192&pay_recode(No)	0.7879	0.2121
4	management_clean_recode(Not Available)&water_tech_category_recode(, Mechanized Pump)&pay_recode(No)&is_urban(True)	0.7538	0.2462
5	management_clean_recode(Not Available)&water_tech_category_recode(, Mechanized Pump)&pay_recode(Yes)&distance_to_primary_road>=6,379	0.7148	0.2852
6	management_clean_recode(Not Available)&water_tech_category_recode(Hand Pump)&served_population<1192&pay_recode(No)	0.6774	0.3226
	management_clean_recode(Not Available)&water_tech_category_recode(Hand Pump)&served_population<1192&pay_recode(Yes)	0.4886	0.5114
	management_clean_recode(Not Available)&water_tech_category_recode(Hand Pump)&served_population>=1192&pay_recode(Yes)	0.4306	0.5694
	^&served_population>=676&pay_recode(No)&is_urban(False)&water_tech_category_recode(Mechanized Pump)&management_clean_recode(Community Management, Direct Government Operation)	0.3897	0.6103
	^&served_population>=676&pay_recode(No)&is_urban(True)&management_clean_recode(Direct Government Operation, Community Management)	0.3351	0.6649
	management_clean_recode(Not Available)&water_tech_category_recode(, Mechanized Pump)&pay_recode(Yes)&distance_to_primary_road<6,379	0.3272	0.6728
	management_clean_recode(Direct Government Operation, Other Institution, Community Management)&served_population>=676&pay_recode(No)&is_urban(False)&water_tech_category_recode(, Hand Pump)	0.3259	0.6741
	^&served_population<676&water_tech_category_recode(, Mechanized Pump)&management_clean_recode(Community Management, Direct Government Operation)&pay_recode(No)	0.3127	0.6873
	^&served_population>=676&pay_recode(No)&is_urban(False)&water_tech_category_recode(Mechanized Pump)&management_clean_recode(Other Institution)	0.2998	0.7002
	^&served_population<676&water_tech_category_recode(Hand Pump)&management_clean_recode(Community Management, Other Institution)&is_urban(True)	0.2656	0.7344
	^&served_population>=676&pay_recode(Yes)&management_clean_recode(Community Management, Direct Government Operation)	0.2575	0.7425
	^&served_population<676&water_tech_category_recode(, Mechanized Pump)&management_clean_recode(Other Institution)&pay_recode(No)&is_urban(False)	0.2573	0.7427
	^&served_population<676&water_tech_category_recode(Hand Pump)&management_clean_recode(Direct Government Operation)&is_urban(True)	0.2535	0.7465
	^&served_population<676&water_tech_category_recode(Hand Pump)&management_clean_recode(Community Management, Other Institution)&is_urban(False)&served_population>=243	0.2416	0.7584
	^&served_population<676&water_tech_category_recode(, Mechanized Pump)&management_clean_recode(Community Management, Direct Government Operation)&pay_recode(Yes)	0.2074	0.7926
	^&served_population<676&water_tech_category_recode(Hand Pump)&management_clean_recode(Direct Government Operation)&is_urban(False)&pay_recode(No)&served_population>=381	0.2039	0.7961
	^&served_population<676&water_tech_category_recode(Hand Pump)&management_clean_recode(Direct Government Operation)&is_urban(False)&pay_recode(Yes)&distance_to_secondary_road<22,183	0.1988	0.8012
	^&served_population<676&water_tech_category_recode(Hand Pump)&management_clean_recode(Community Management, Other Institution)&is_urban(False)&served_population<243	0.1961	0.8039
	^&served_population>=676&pay_recode(No)&is_urban(True)&management_clean_recode(Other Institution)	0.1948	0.8052
	^&served_population<676&water_tech_category_recode(, Mechanized Pump)&management_clean_recode(Other Institution)&pay_recode(No)&is_urban(True)	0.1566	0.8434
	^&served_population<676&water_tech_category_recode(, Mechanized Pump)&management_clean_recode(Other Institution)&pay_recode(Yes)&is_urban(False)	0.1522	0.8478
	^&served_population>=676&pay_recode(Yes)&management_clean_recode(Other Institution)	0.1516	0.8484
	^&served_population<676&water_tech_category_recode(Hand Pump)&management_clean_recode(Direct Government Operation)&is_urban(False)&pay_recode(No)&served_population<381	0.1333	0.8667
	^&served_population<676&water_tech_category_recode(, Mechanized Pump)&management_clean_recode(Other Institution)&pay_recode(Yes)&is_urban(True)	0.0219	0.9781

Figure 12: Decision Tree Leaf Report of Splitting Rules

In summary, the generalized rules for identifying non-functional groups can be derived as followed. A water point is likely to be nonfunctional if it:

- Has no management structure and is far from primary road, regardless of payment scheme. Within this group,
  - The mechanized pumps perform more poorly than handpump.
  - If the handpump serves more than ~1200 people without payment scheme, it will perform as poor as the mechanized pump. Meaning that payment scheme may be of some help in relatively larger communities equipped with handpumps, otherwise it is not very useful.
- Is with management structure but far from secondary road.

Lastly, the overall fit and performance of the model is observed in **Figure 13** below. We can see that the two models are comparable in terms of both Misclassification rate and Confusion Matrix. Despite performing slightly worse in the true-positive rate (lower by ~0.005 or 0.5%), the decision tree model is more user friendly and can be used to draw conclusions faster with multi-variable grouping.



A – Decision Tree

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1739	0.1683	0.1779	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2815	0.2735	0.2873	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5533	0.5574	0.5508	∑ -Log(ρ[j])/n
RASE	0.4297	0.4319	0.4282	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3694	0.3721	0.3682	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2549	0.2601	0.2514	∑ (ρ[j]≠ρMax)/n
N	57759	8203	8233	n

Confusion Matrix

	Actual status_recode	Predicted Count Nonfunctional	Predicted Count Functional	Predicted Rate Nonfunctional	Predicted Rate Functional
Training	Nonfunctional	12193	10455	0.538	0.462
Training	Functional	4267	30844	0.122	0.878
Validation	Nonfunctional	1700	1525	0.527	0.473
Validation	Functional	609	4369	0.122	0.878
Test	Nonfunctional	1756	1479	0.543	0.457
Test	Functional	591	4407	0.118	0.882

B – Reduced Log Regression

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1613	0.1554	0.1651	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2632	0.2546	0.2689	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5616	0.5656	0.5594	∑ -Log(ρ[j])/n
RASE	0.4332	0.4354	0.4320	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3753	0.3778	0.3747	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2564	0.2601	0.2531	∑ (ρ[j]≠ρMax)/n
N	57077	8100	8136	n

Lack Of Fit

Parameter Estimates

Effect Likelihood Ratio Tests

Confusion Matrix

	Actual status_recode	Predicted Count Nonfunctional	Predicted Count Functional	Predicted Rate Nonfunctional	Predicted Rate Functional
Training	Nonfunctional	12142	10231	0.543	0.457
Training	Functional	4405	30299	0.127	0.873
Validation	Nonfunctional	1691	1485	0.532	0.468
Validation	Functional	622	4302	0.126	0.874
Test	Nonfunctional	1746	1451	0.546	0.454
Test	Functional	608	4331	0.123	0.871

Figure 13: Comparison of Decision Tree and Logistic Regression Model Performance

5. CONCLUSTION AND RECOMMENDATIONS

From the study above, some predictors of nonfunctional waterpoints can be identified, most notably the presence of management structure, whether the pump is manual or mechanized, distance to primary roads, served population and the presence of payment schemes. The decision tree model is more user friendly and can be deployed easily by non-technical audience with specific targeted high-risk groups (high proportion of nonfunctional water points) with abovementioned rules.

However, in overall both models still suffer from the lack of fit and relatively low true-positive rate. This can be associated with the removal of many potentially useful variables such as install_year and geographical locations, which respectively carry information on the age of facility and natural conditions and/or accessibility to technical resources of the community. Recommendations can be listed as followed:

- Consider grouping the geographical locations into meaningful groups by obtaining more understanding of the country and assign the grouping to the rows based on their administrative or latitudinal/longitudinal locations. The groupings can be based on socio-economic (degree of economic development) or natural factors (mountainous vs coastal areas). This can potentially result in good predictors of nonfunctionality as waterpoints operation and maintenance can depend on the economic and natural conditions that the communities are living in.

- Consider limiting the analysis on the rows where install_year is available. At a cost of less datapoints, this may give another meaningful predictor of facility age (Foster, 2013)

REFERENCE

Liddle, Elisabeth & Fenner, Richard. (2017). Water point failure in sub-Saharan Africa: The value of a systems thinking approach. Waterlines. 36. 140-166. 10.3362/1756-3488.16-00022.

Foster, t. (2013) 'Predictors of sustainability for community-managed handpumps in sub-Saharan Africa: evidence from Liberia, Sierra Leone, and uganda', Environment, Science and Technology 47(21): 12037–46 <http://dx.doi.org/10.1021/es402086n>.