The Data Behind the Tanzanian Water Crisis

Nick Kachanyuk, M.S. Data Science Candidate

August 11, 2021

Willamette University

Atkinson Graduate School of Management

Agenda

- → Project Research Questions
- → Background
- → Data
- → Methods
- → Results
- → Variables of importance
- → Recommendations
- → Takeaways
- → Q&A

Project Research Questions

- What machine learning algorithm is suitable for classifying water well functionality?
- What are the important variables for the machine learning approach?
- What are some general things that differentiate a working water well from wells that are non-functional and/or need repair?

Background

- According to Water.org records, about 4 million people lack access to safe water resources in Tanzania [1].
- In 2016, Water.org found that Tanzania is eligible for a “water credit solution”.
  - Lending solutions to households, water companies, local government, etc.
- Currently there are no known guidelines established to tackle the issue
- My project attempts to investigate available data and provide actionable guidelines for stakeholders

Data

- Two sources of data were used:
  - DrivenData [2]
    - 39 features + target variable
    - 31 categorical, 8 numeric
    - Multiclassification problem
      - Functional, functional needs repair, non-functional
  - 2012 Tanzania census data
    - 7 numeric variables on region specific demographics (population, average household size, unemployment rate, region area, etc.)
    - 2 feature engineered variables
      - Population density
      - Well strain

Methods

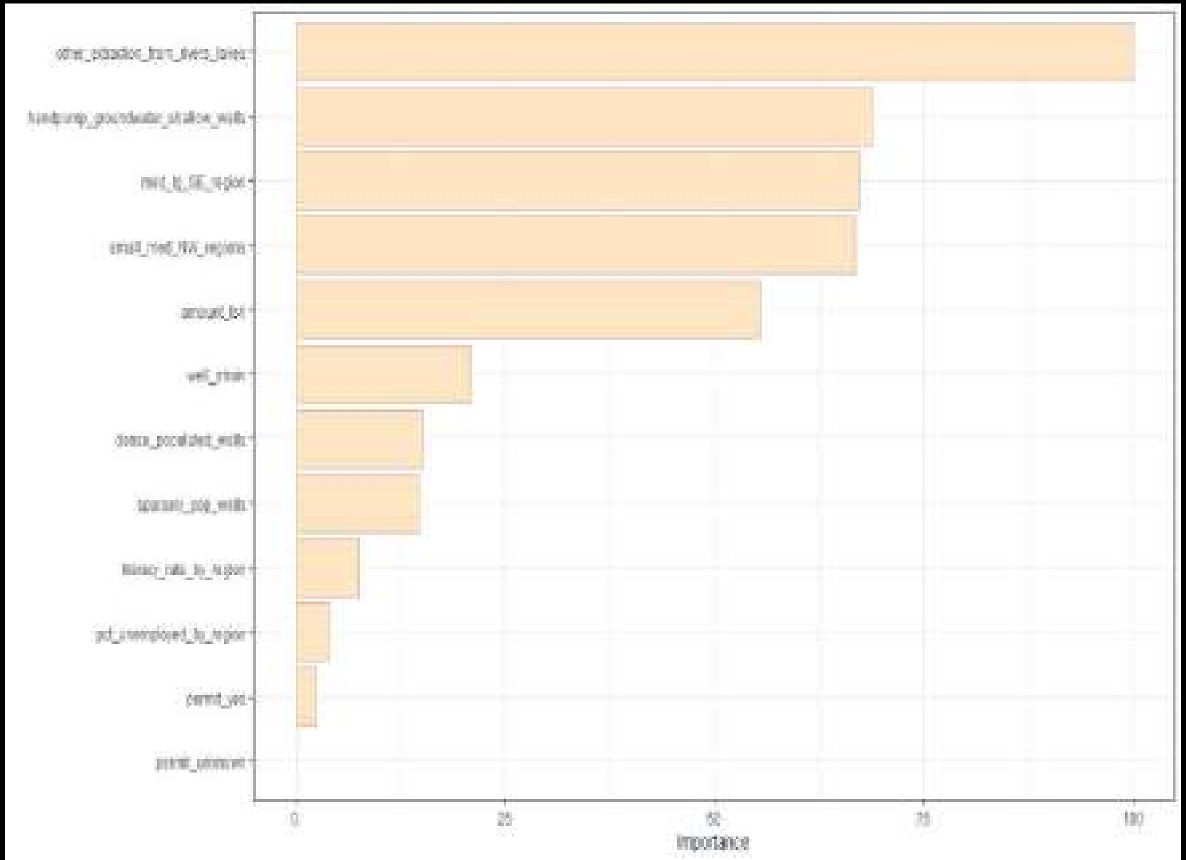
- Imputation of missing data (median, mean)
    - Data was not missing at random
- Principal Component Analysis
    - Dimensionality reduction
    - Many categorical variables; even more dummy variables
- 4 XGBoost models
    - Handles outliers well
    - Less prone to overfitting
    - Known for good model performance

Results

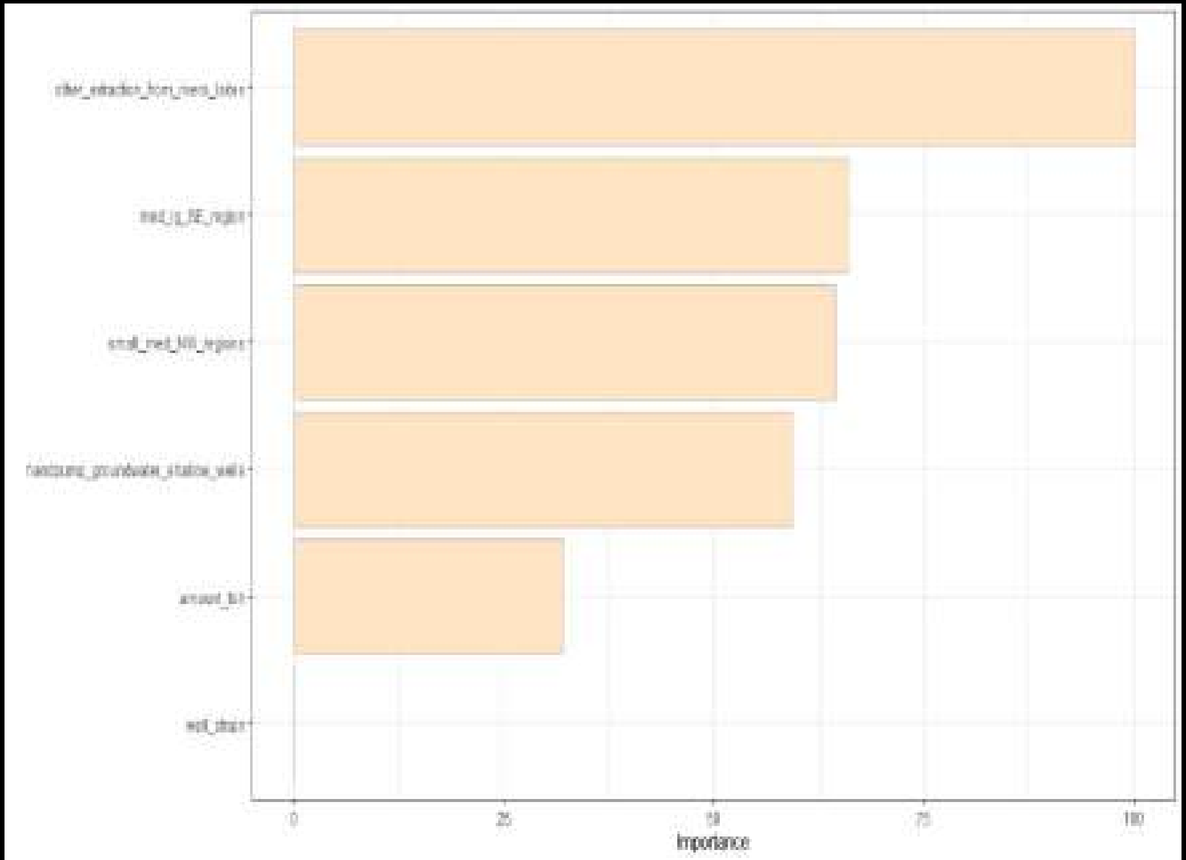
- Four final models were selected:
    - Baseline model:
        - 12 predictors, 3 target classes
        - Kappa: 0.599
    - Model 2:
        - 6 predictors, 3 target classes
        - Kappa: 0.600
    - Model 3:
        - 12 predictors, 2 target classes
        - Kappa: 0.598
    - Model 4:
        - 6 predictors, 2 target classes
        - Kappa: 0.594

Variables of Importance

Models w/ 12 predictors

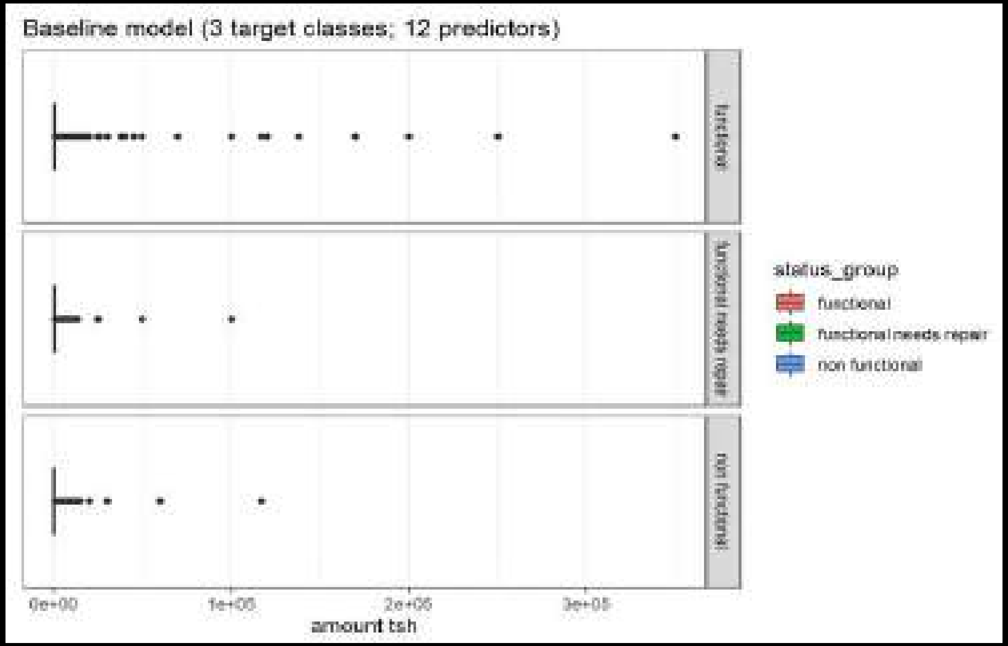


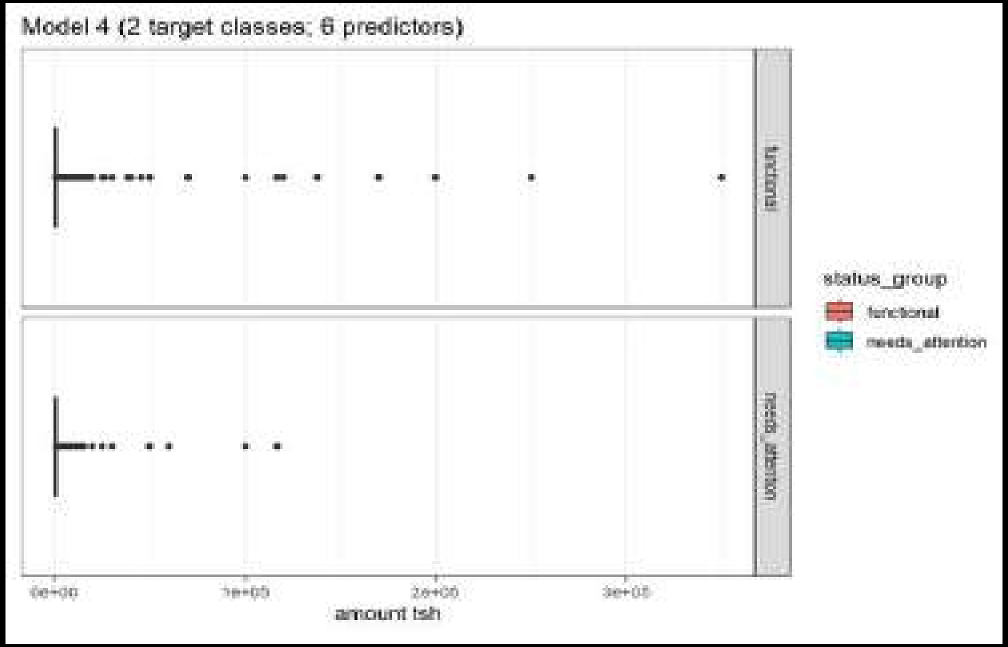
Models w/ 6 predictors

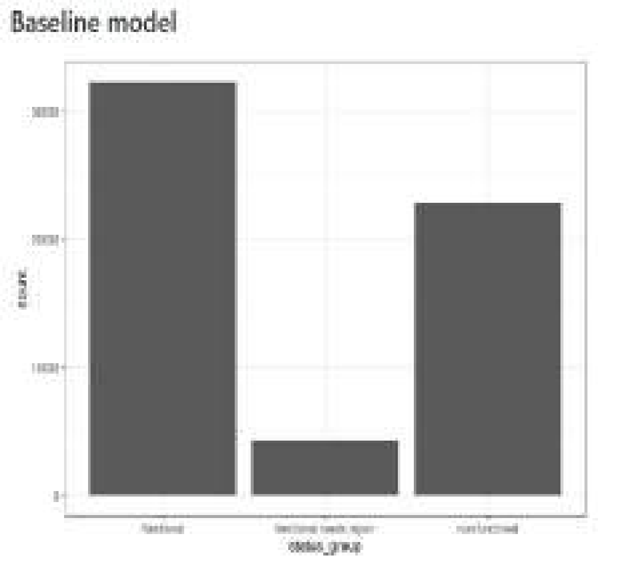


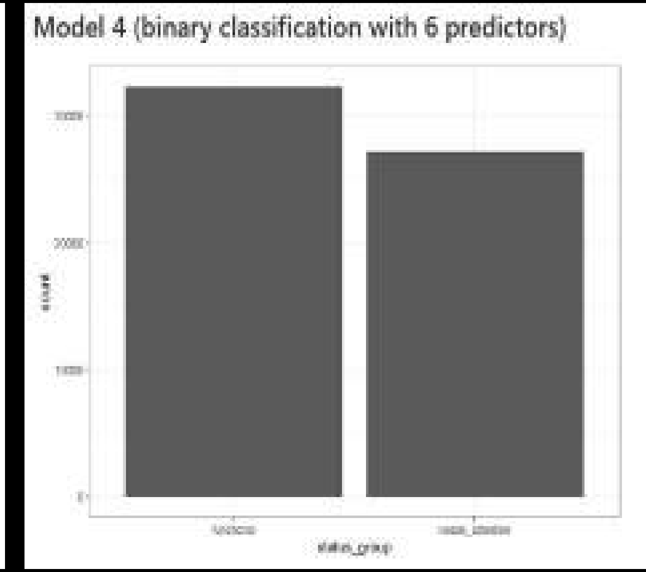
Recommendations

- Use model 4 (or something similar to it)!
    - Comparable Kappa score to other models
    - Only uses 6 predictors (more succinct/interpretable)
    - Target classes are well defined and more balanced









Takeaways

- Functional wells:
    - Don’t rely on handpump or “other” extraction methods
    - Tend to be located more in southern and eastern regions of Tanzania
    - Have more water volume available within the well/waterpoint
    - Experience less strain
- Needs attention wells:
    - Rely on handpump or “other” extraction
    - Tend to be located more in northern and western regions of Tanzania
    - Have less water volume available within the well/waterpoint
    - Experience higher well strain

Thank you

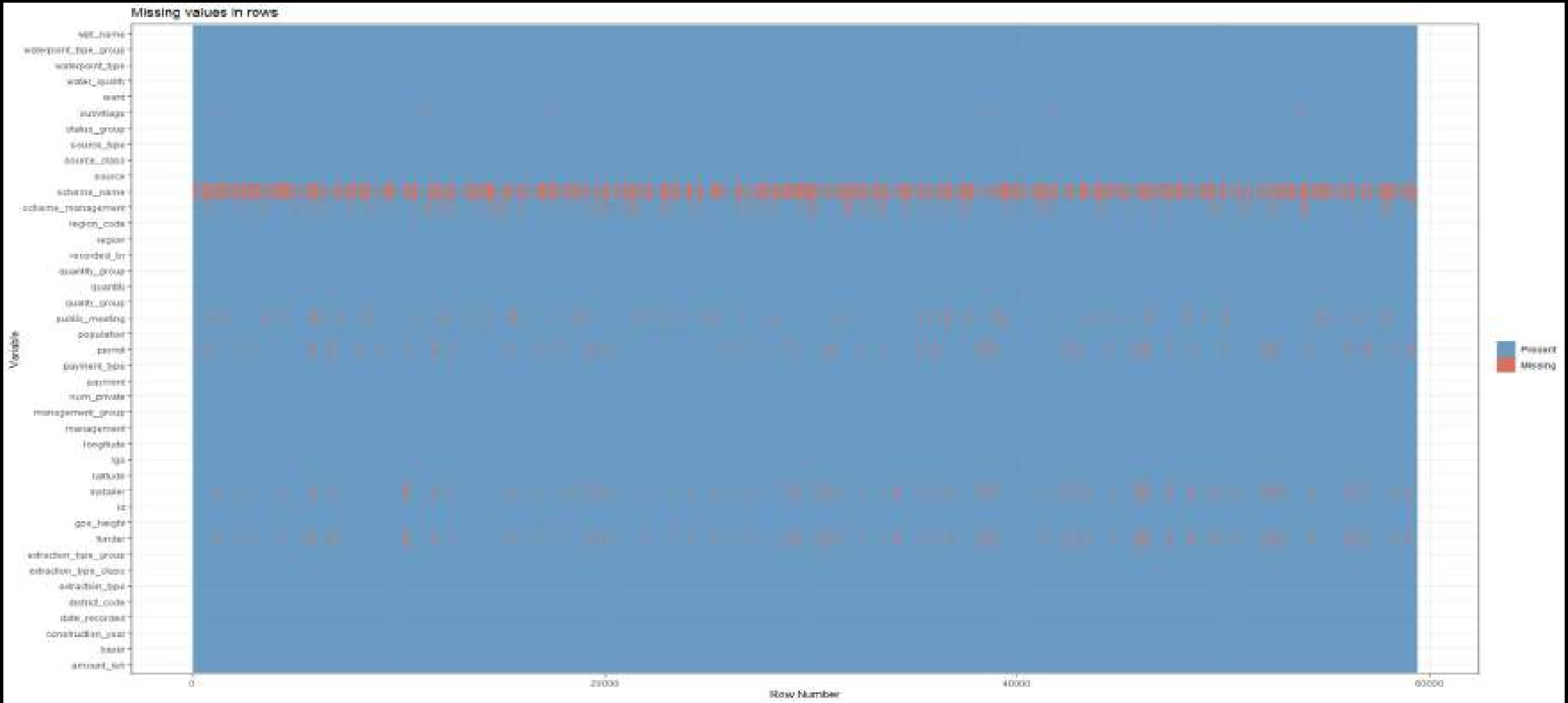
Nick Kachanyuk | nick-kachanyuk-website.netlify.app

nkachanyuk123@gmail.com | github.com/nickkachanyuk

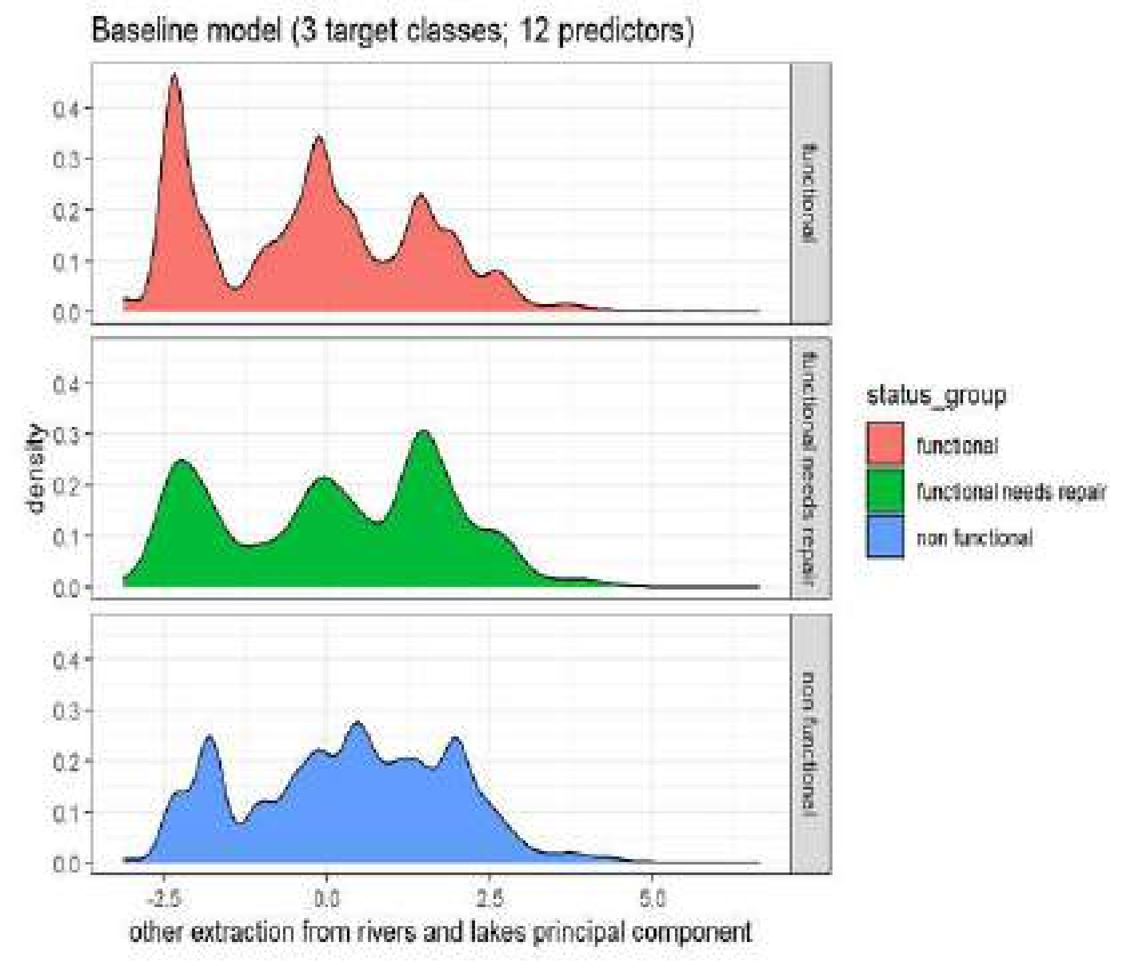
References

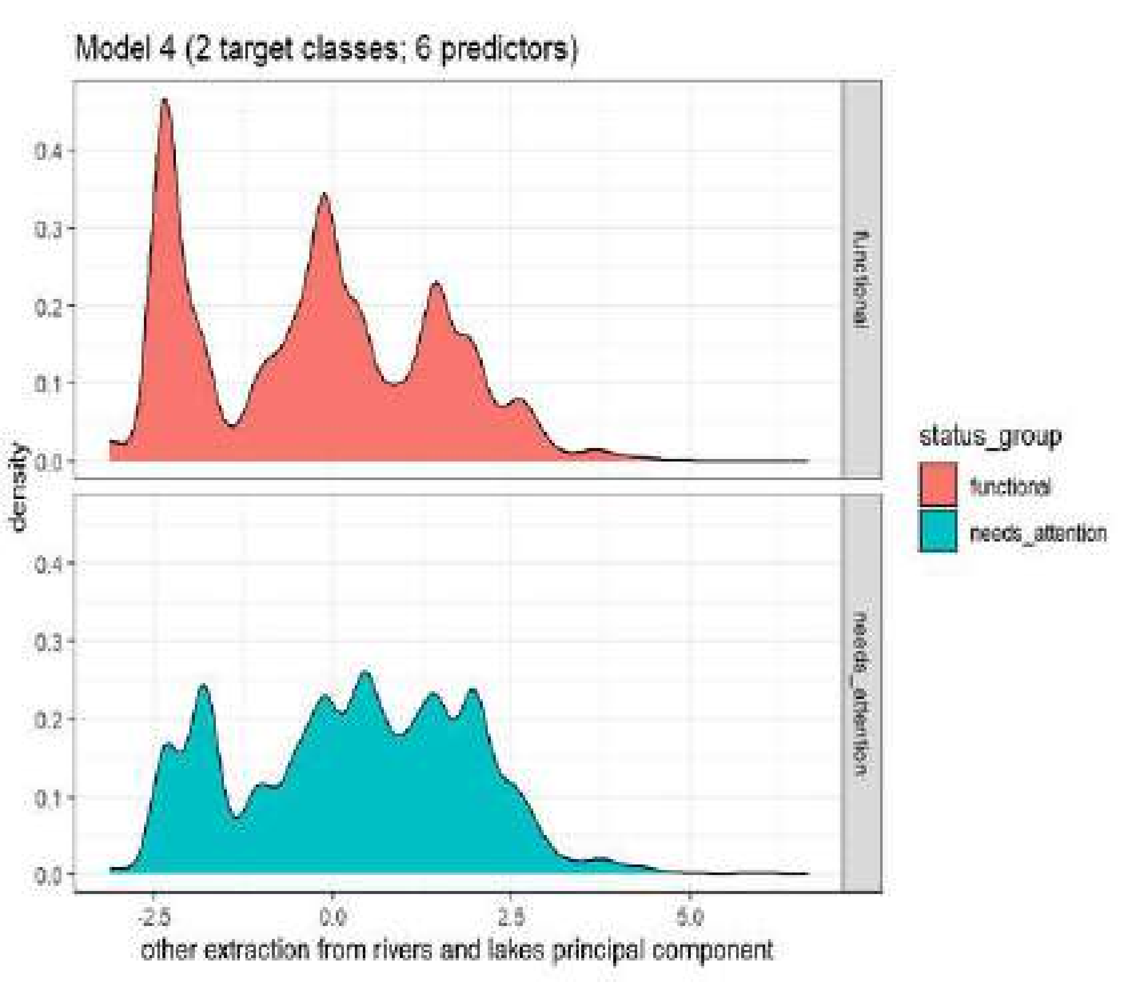
- [1] *Tanzania's water crisis - Tanzania's water in 2021*. Water.org. (n.d.). https://water.org/our-impact/where-we-work/tanzania/.
- [2] DrivenData. (n.d.). *Pump it Up: Data mining the water table*. DrivenData. https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25/.

Backup slides – Missing data graph

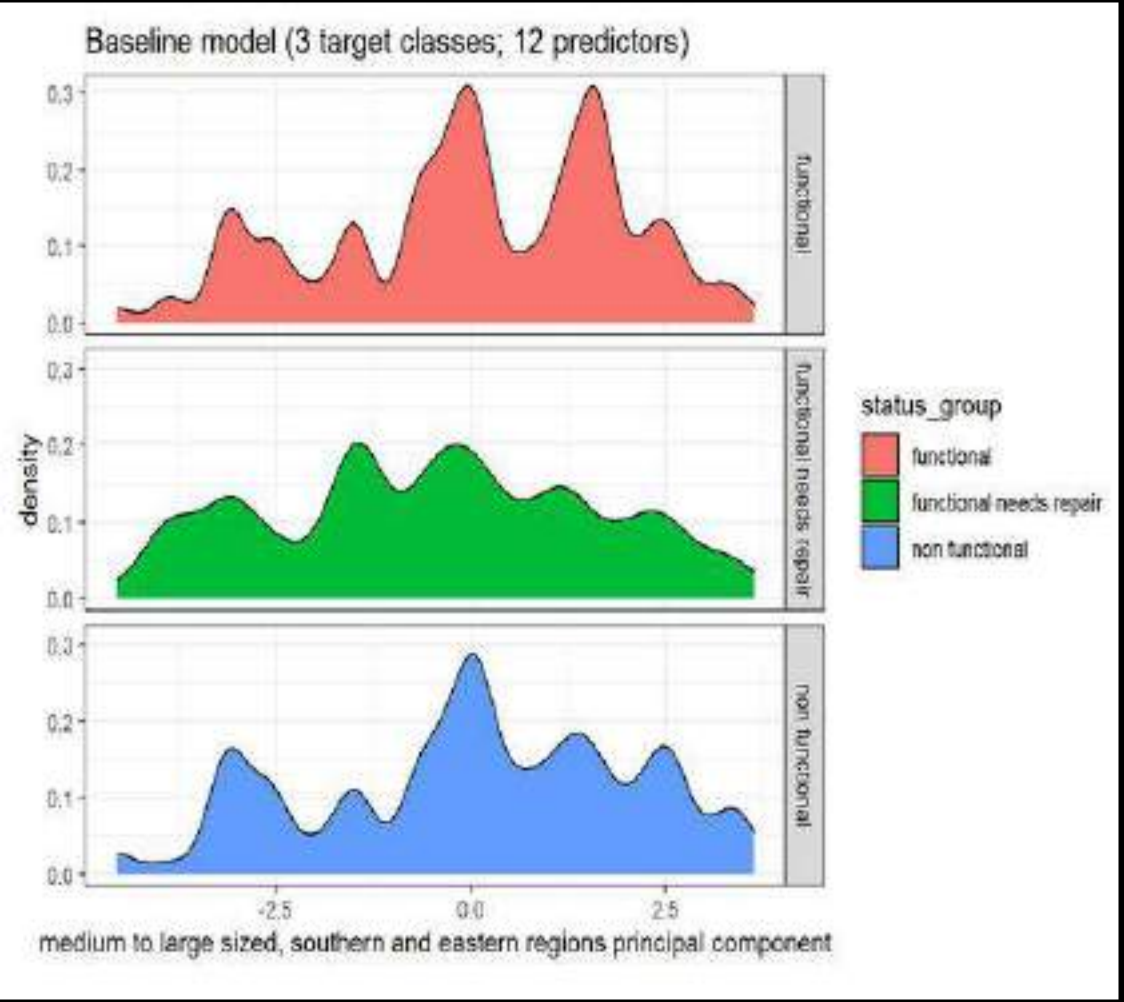


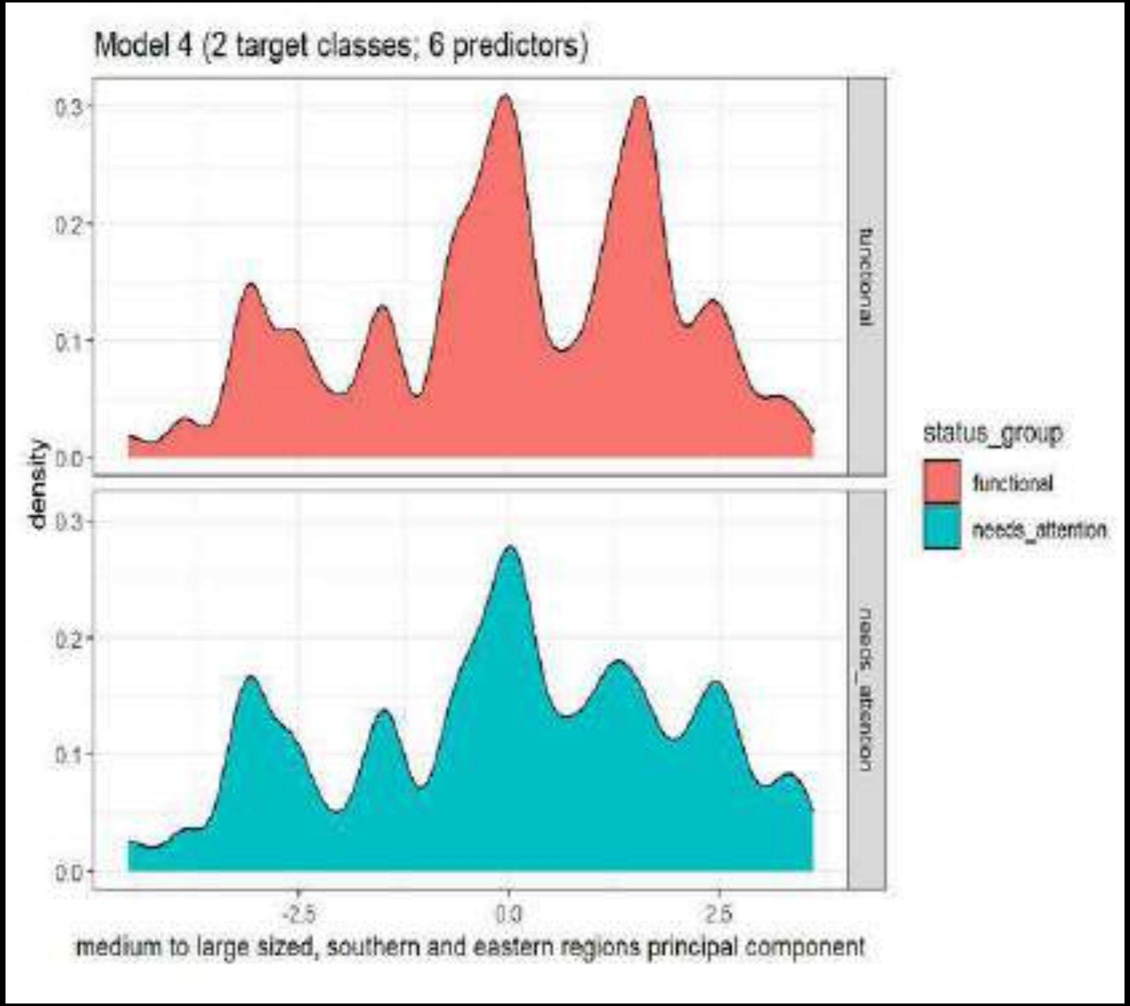
Backup slides – Other extraction from rivers , lakes



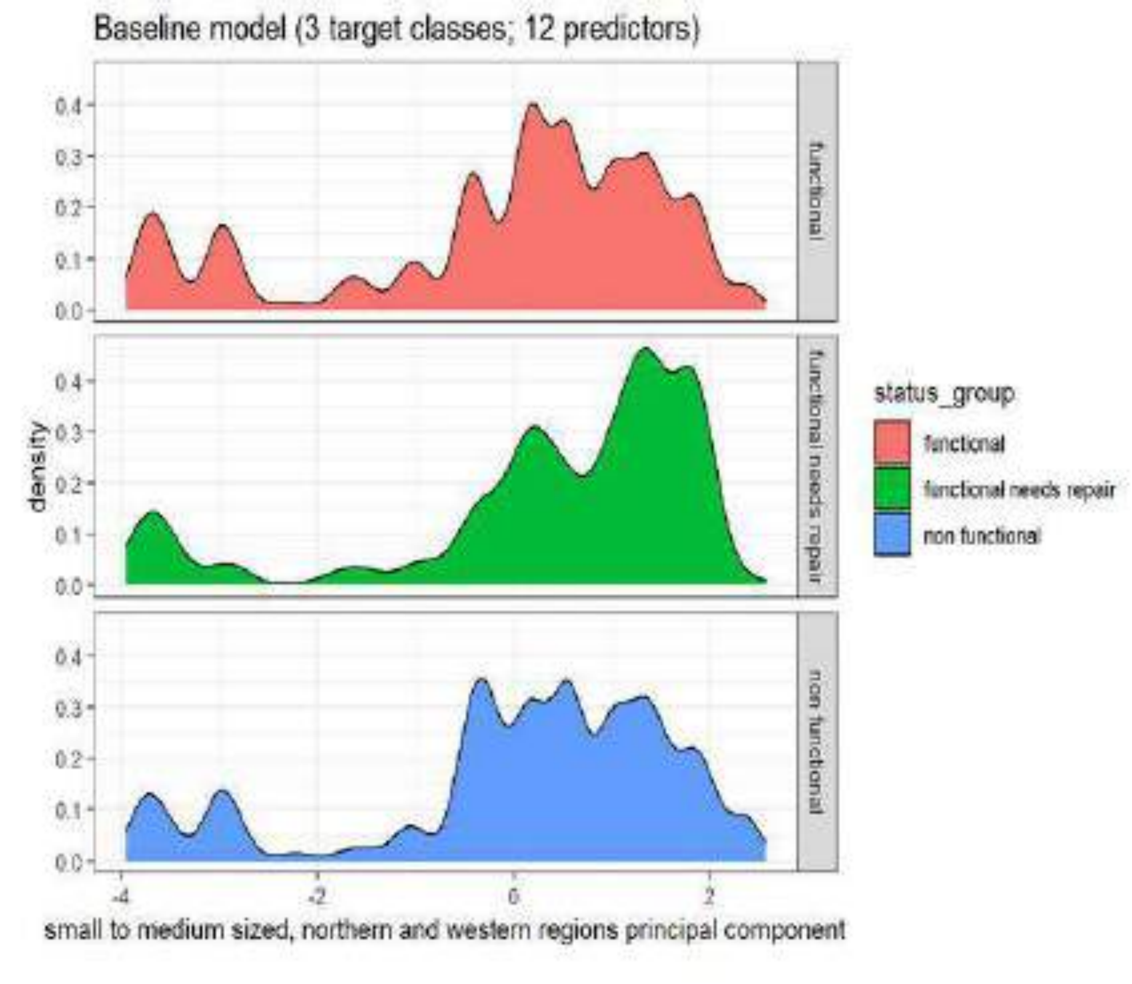


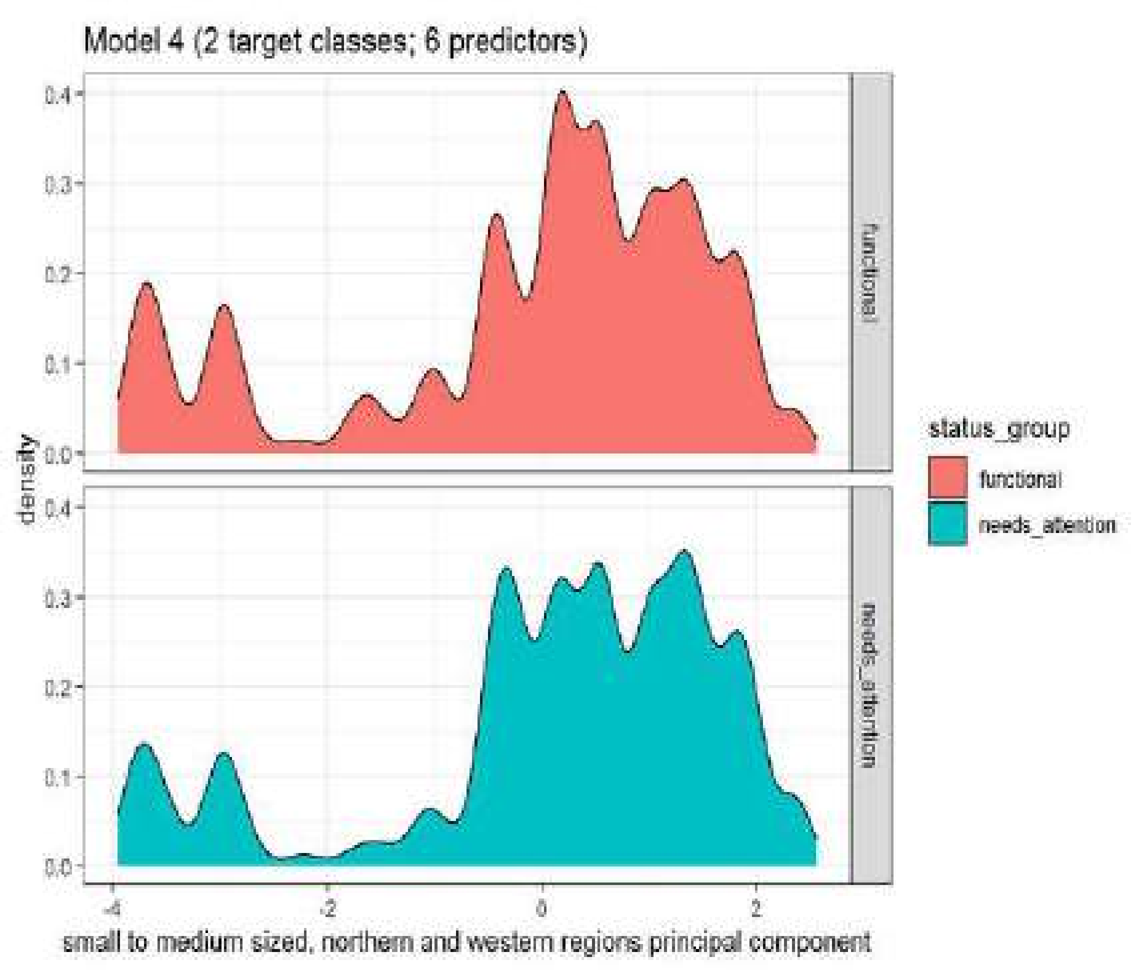
Backup slides – Medium to large sized, southern and eastern regions principal component variable



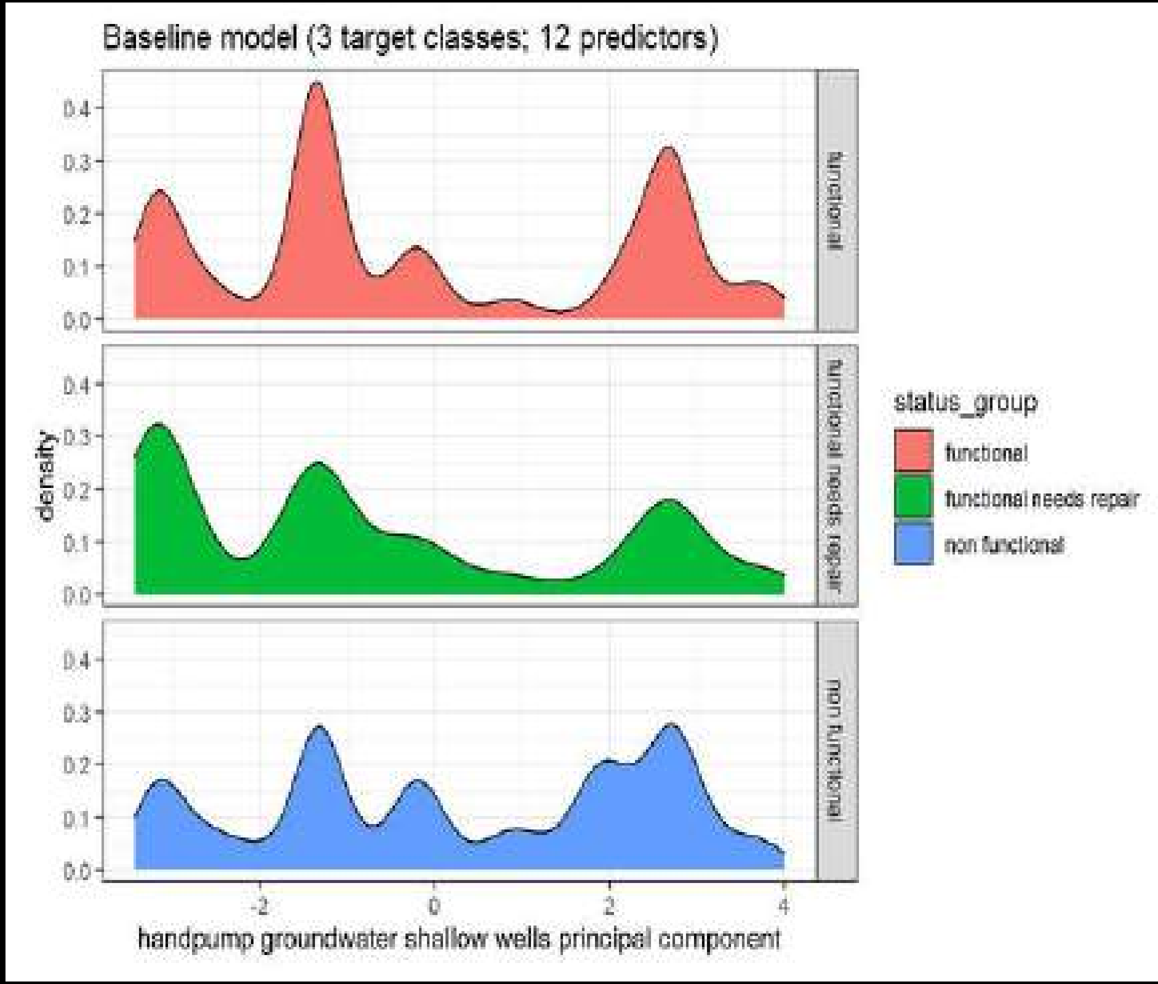


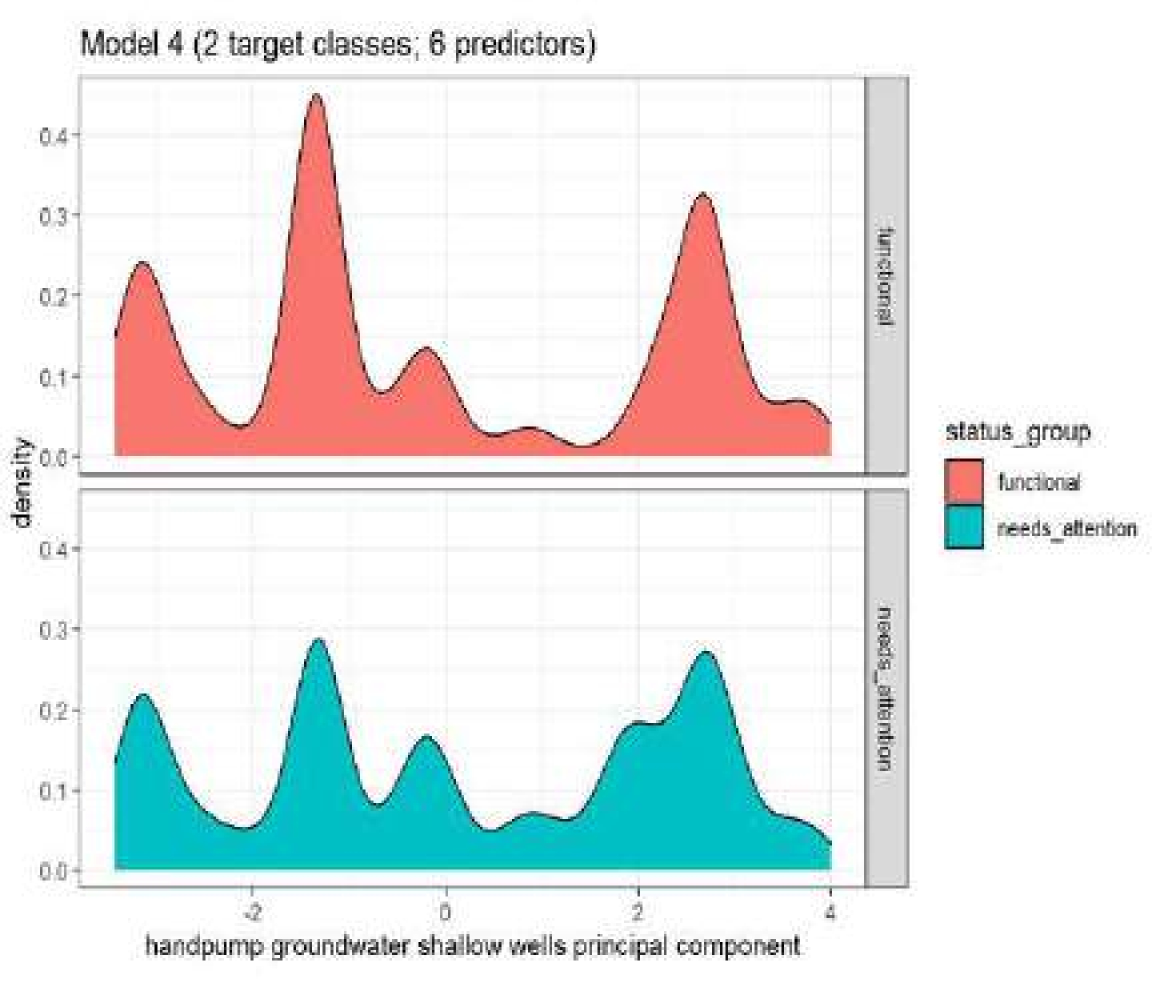
Backup slides – Small to medium sized, northern and western regions principal component variable



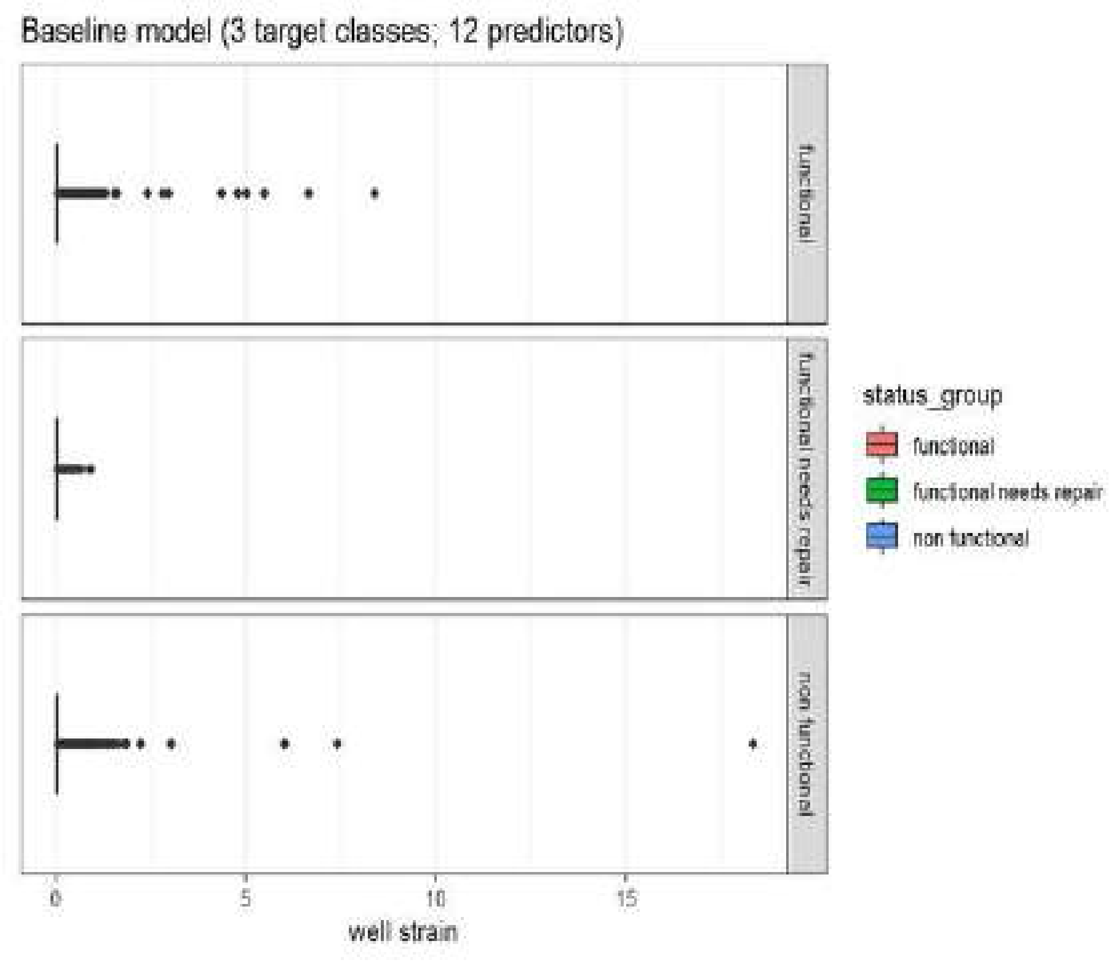


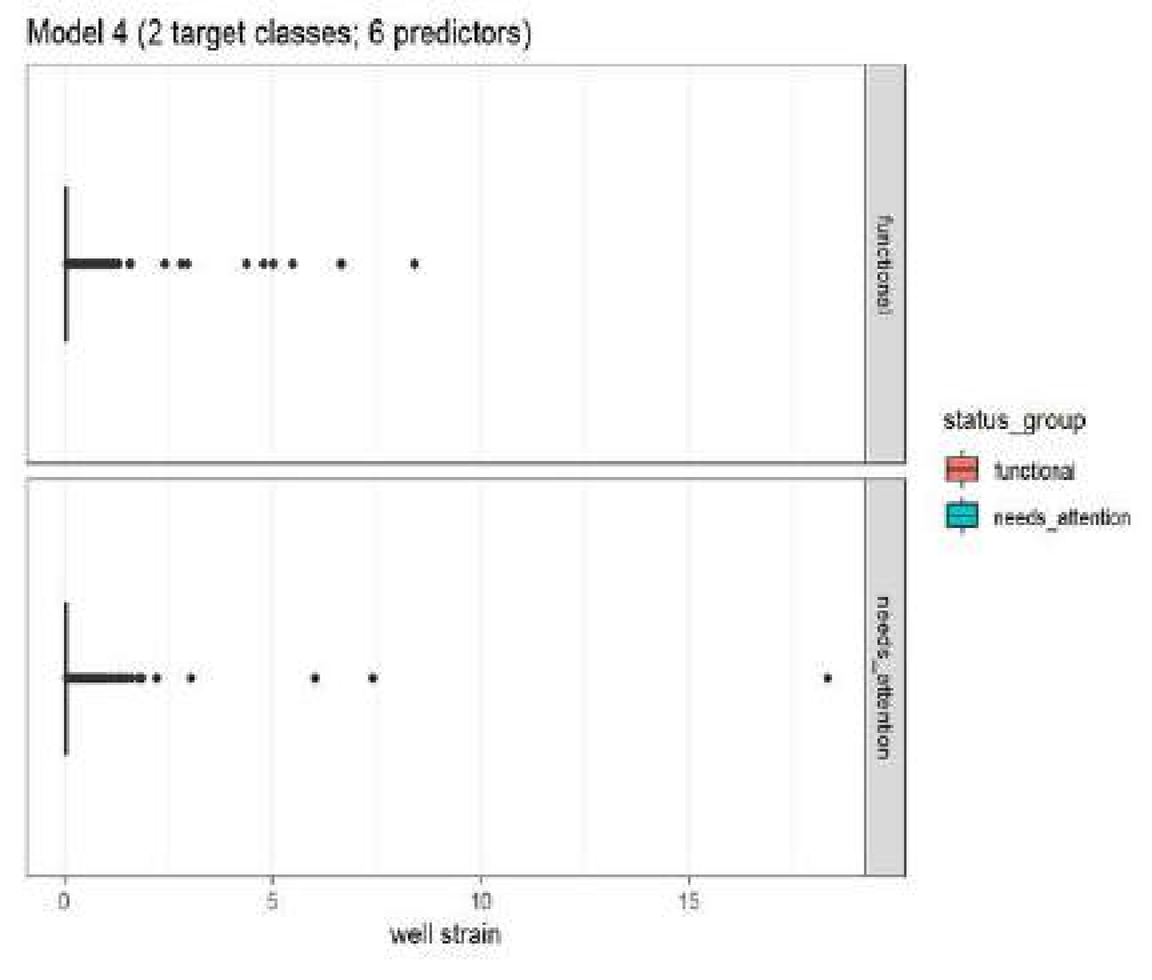
Backup slides – Handpump groundwater shallow well principal component variable





Backup slides – Well strain variable





Backup slides – Well strain variable formula

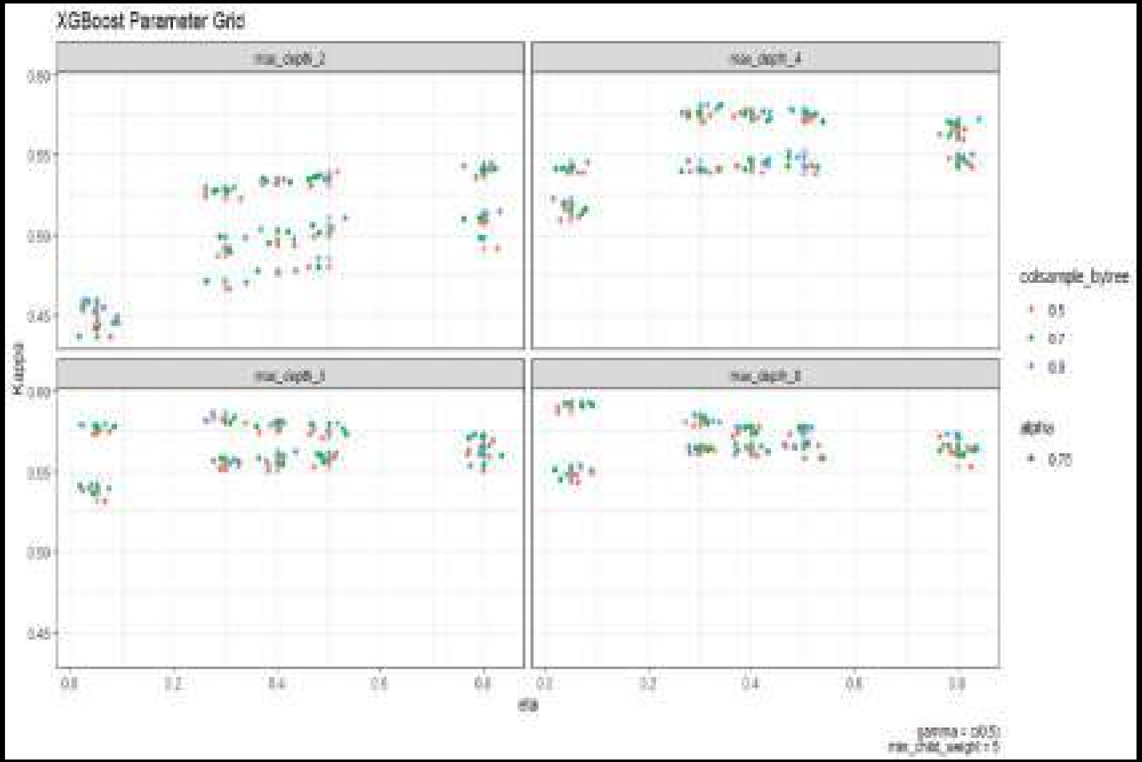
- Formula:

  well_strain = ((population/total_region_pop)∗region_pop_density)

  - where population is the number of people living around the well
  - total_region_pop is the total number of people living in a given region of Tanzania
  - region_pop_density is total_region_pop/region_area_sq_mi

Backup slides – Model 1 (baseline model)

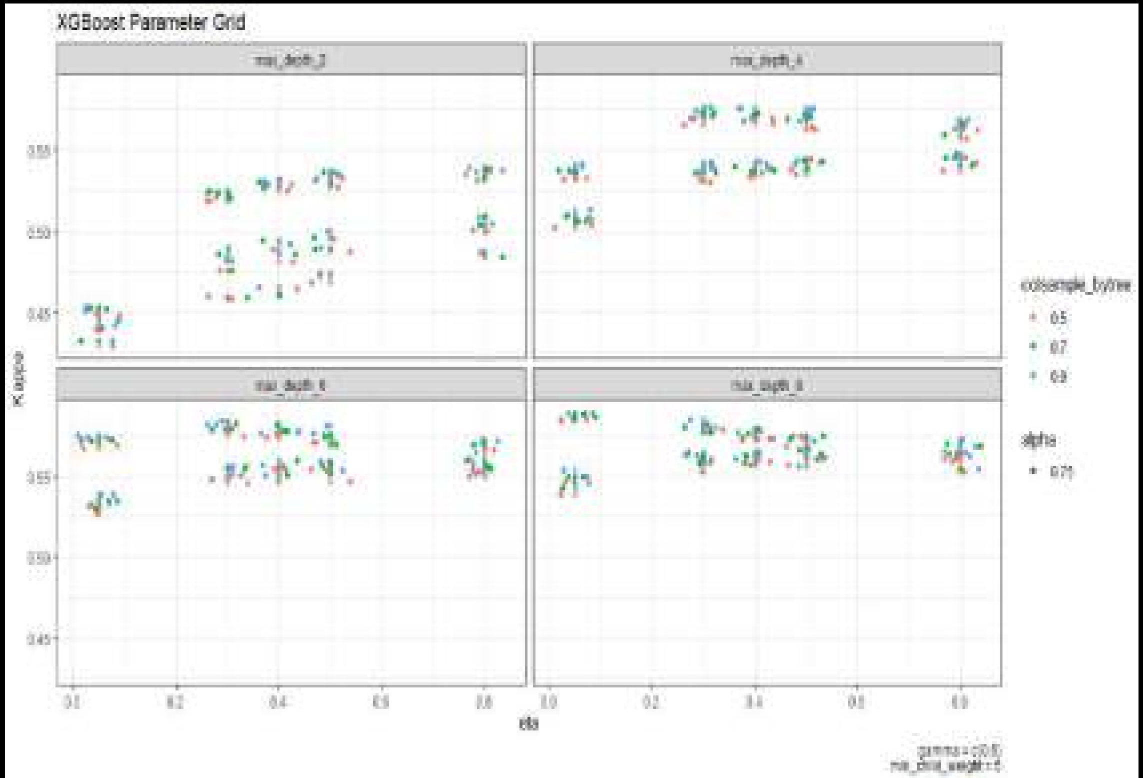
```
Confusion Matrix and Statistics

                         Reference
Prediction                functional functional needs repair non functional
  functional                    7101                     608           1366
  functional needs repair        206                     307             72
  non functional                 764                     142           4284

Overall Statistics

               Accuracy : 0.7873
                 95% CI : (0.7807, 0.7939)
    No Information Rate : 0.5435
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.599

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: functional Class: functional needs repair Class: non functional
Sensitivity                     0.8798                        0.29044                0.7487
Specificity                     0.7088                        0.97984                0.9007
Pos Pred Value                  0.7825                        0.52479                0.8254
Neg Pred Value                  0.8320                        0.94742                0.8511
Prevalence                      0.5435                        0.07118                0.3853
Detection Rate                  0.4782                        0.02067                0.2885
Detection Prevalence            0.6111                        0.03939                0.3495
Balanced Accuracy               0.7943                        0.63514                0.8247
```



	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
61	1000	8	0.05	0	0.9	5	0.5

Backup slides – Model 2 (6 predictors, 3 target classes)

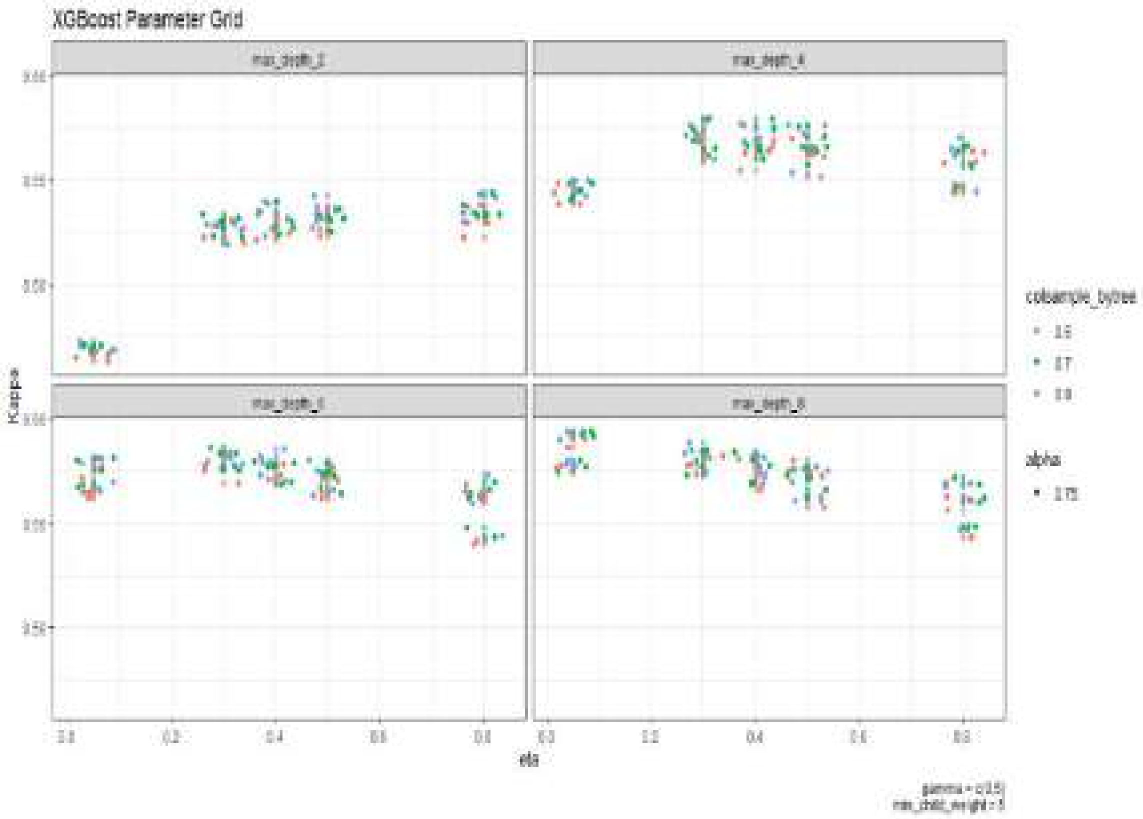
```
Confusion Matrix and Statistics

                        Reference
Prediction               functional functional needs repair non functional
  functional                   7160                     626           1398
  functional needs repair       192                     296             68
  non functional                719                     135           4256

Overall Statistics

               Accuracy : 0.7887
                 95% CI : (0.782, 0.7952)
    No Information Rate : 0.5435
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6003

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: functional Class: functional needs repair Class: non functional
Sensitivity                     0.8871                        0.28004                0.7438
Specificity                     0.7014                        0.98115                0.9064
Pos Pred Value                  0.7796                        0.53237                0.8329
Neg Pred Value                  0.8392                        0.94676                0.8495
Prevalence                      0.5435                        0.07118                0.3853
Detection Rate                  0.4822                        0.01993                0.2866
Detection Prevalence            0.6185                        0.03744                0.3441
Balanced Accuracy               0.7943                        0.63059                0.8251
```



	nrounds <dbl>	max_depth <dbl>	eta <dbl>	gamma <dbl>	colsample_bytree <dbl>	min_child_weight <dbl>	subsample <dbl>
62	1000	8	0.05	0	0.9	5	0.75

Backup slides – Model 3 (12 predictors, 2 target classes)

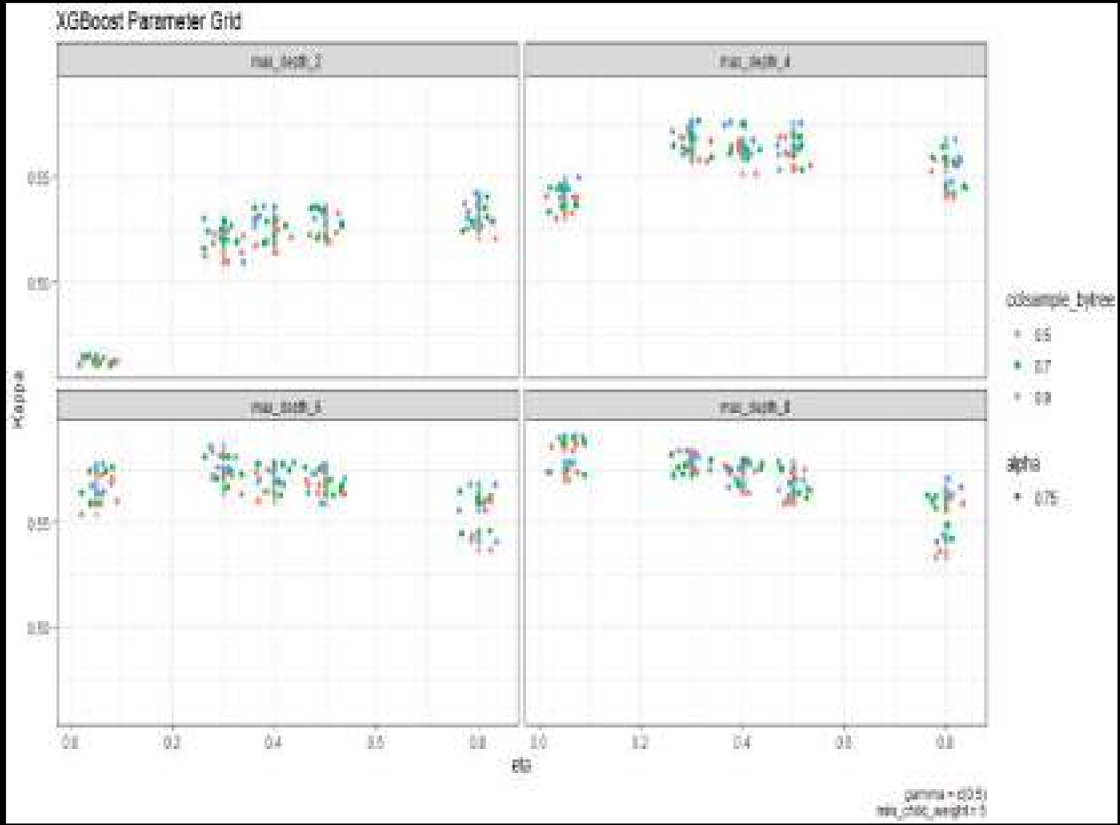
```
Confusion Matrix and Statistics

                 Reference
Prediction        functional needs_attention
  functional            6916            1788
  needs_attention       1155            4991

               Accuracy : 0.8018
                 95% CI : (0.7953, 0.8082)
    No Information Rate : 0.5435
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5976

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8569
            Specificity : 0.7362
         Pos Pred Value : 0.7946
         Neg Pred Value : 0.8121
             Prevalence : 0.5435
         Detection Rate : 0.4657
   Detection Prevalence : 0.5861
      Balanced Accuracy : 0.7966

       'Positive' Class : functional
```



	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
63	1000	8	0.05	0	0.9	5	0.9

Backup slides – Model 4 (6 predictors, 2 target classes)

```
Confusion Matrix and Statistics

                 Reference
Prediction        functional needs_attention
  functional            6915            1813
  needs_attention       1156            4966

               Accuracy : 0.8001
                 95% CI : (0.7935, 0.8065)
    No Information Rate : 0.5435
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5939

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8568
            Specificity : 0.7326
         Pos Pred Value : 0.7923
         Neg Pred Value : 0.8112
             Prevalence : 0.5435
         Detection Rate : 0.4657
   Detection Prevalence : 0.5877
      Balanced Accuracy : 0.7947

       'Positive' Class : functional
```



	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
63	1000	8	0.05	0	0.9	5	0.9