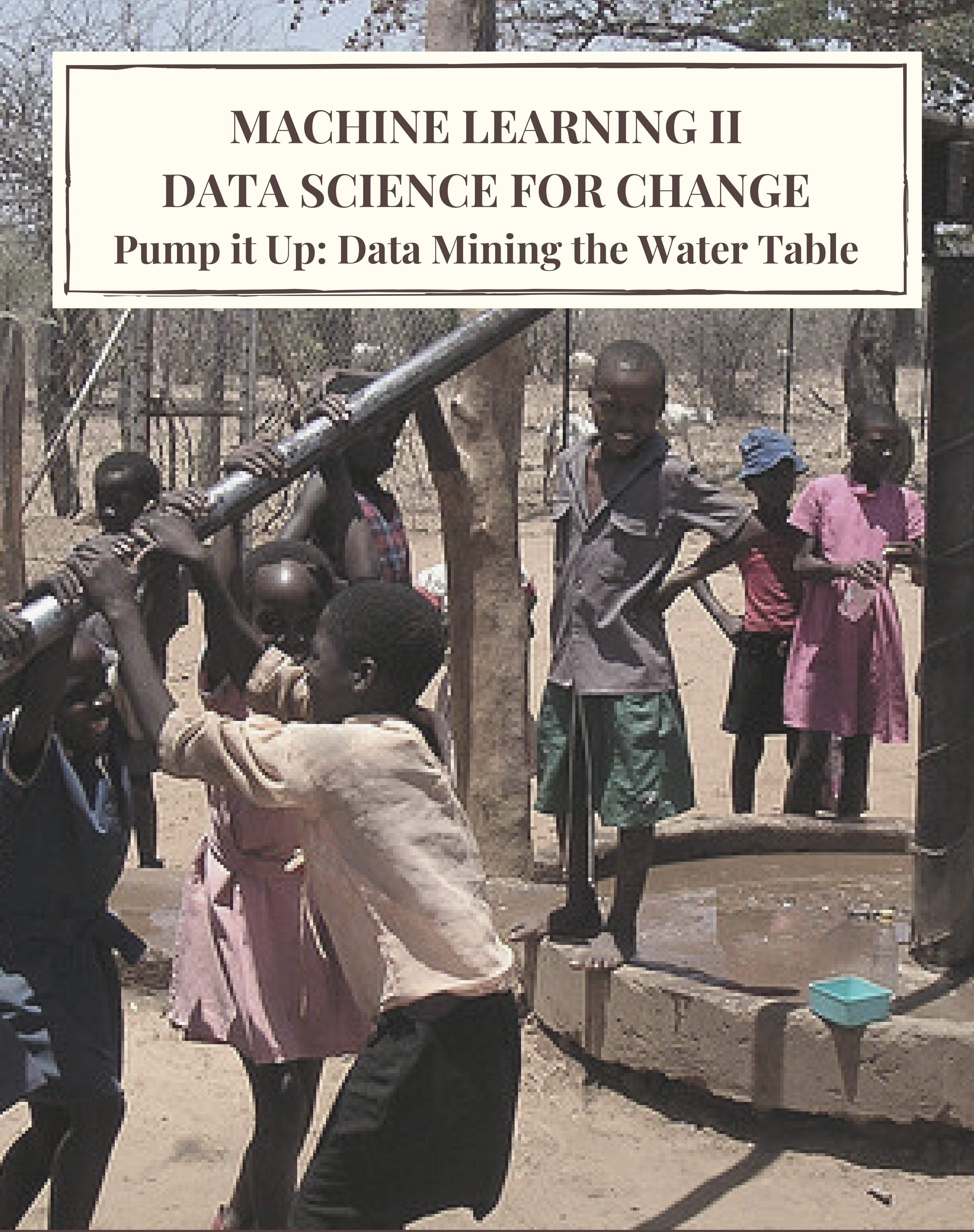
MACHINE LEARNING II
DATA SCIENCE FOR CHANGE
Pump it Up: Data Mining the Water Table



IE MBD-2 GROUP G
ALEJANDRO BARRERO | PATRICIA CÁRCAMO ACOSTA | AMALA EGGERS
LUIS MIGUEL GALLEGOS | ANNA KOENIG | FRANCISCO VIGO

TABLE
OF CONTENTS

01 OBJECTIVE

02 DATA INSPECTION

03 FEATURE EXPLORATORY ANALYSIS

06 BASELINE MODEL

07 FILTERING METHODS

07 EMBEDDED METHODS

08 MODELLING AND EVALUATION

09 MODEL SELECTION

10 CONCLUSIONS

10 DRIVEN DATA CHALLENGE SUBMISSION

11 CITATIONS

OBJECTIVE

Our challenge is part of the Driven Data Competition "Pump it Up: Data Mining the Water Table" which uses data from Taarifa and the Tanzanian Ministry of Water and is available at this link. The task is to search for data driven solutions to predict which waterpoints will fail in order to improve maintenance operations and ensure clean potable water availability for the communities across Tanzania. According to Tanzania's Ministry of Water, "development comes only when there is plenty of quality water, which is sustainably utilized".

Our task is a multi-class classification problem to predict the state of water pumps in Tanzania into three categories: working, need repair, and fail. This report provides an explanation of our analysis of the features, machine learning process, and evaluation of results.

Taarifa, which means reporting in Swahili, is an open source web API to close citizen feedback loop through data collection, visualization and data mapping. It seeks to address the problems of sanitation, potable water, and waste management in the developing world, particularly Africa. The system works by allowing citizens to report sanitation problems through SMS, webform, email or Twitter. This report is sent directly to authorities who have the power to act on the issue and a geolocation marker is set on the map for quick identification of problem areas.

In this report we will explain our discovery process and the steps we took to reach a final predictive model. This process includes analyzing and filtering the relevant features in our dataset, testing out different models, and selecting the model that provides the best accuracy and explainability.

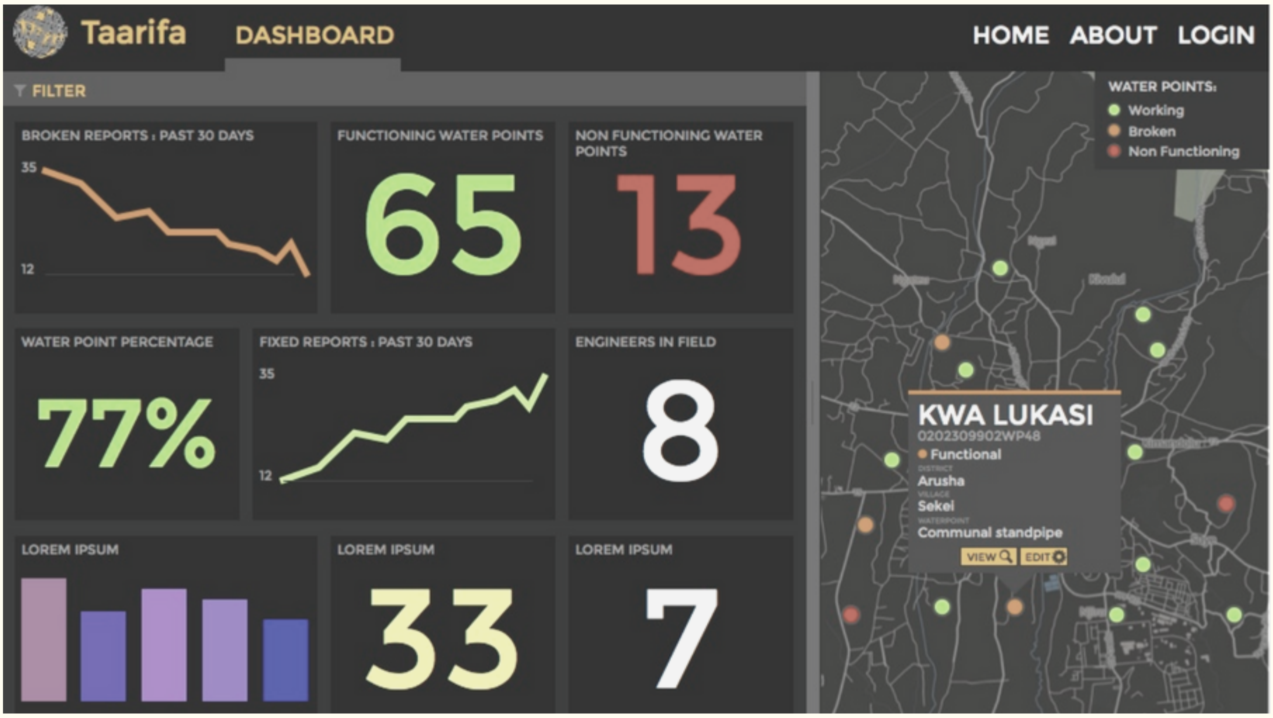


Fig 1 Sample Taarifa dashboard

DATA INSPECTION

The dataset contains 41 features of waterpoints in Tanzania and a total of 74,250 observations out of which 59,400 observations make the training set. The features include:

amount_tsh - Total static head (amount water available to waterpoint)
date_recorded - The date the row was entered
funder - Who funded the well
gps_height - Altitude of the well
installer - Organization that installed the well
longitude - GPS coordinate
latitude - GPS coordinate
wpt_name - Name of the waterpoint if there is one
num_private
basin - Geographic water basin
subvillage - Geographic location
region - Geographic location
region_code - Geographic location (coded)
district_code - Geographic location (coded)
lga - Geographic location
ward - Geographic location
population - Population around the well
public_meeting - True/False
recorded_by - Group entering this row of data
scheme_management - Who operates the waterpoint
scheme_name - Who operates the waterpoint
permit - If the waterpoint is permitted
construction_year - Year the waterpoint was constructed
extraction_type - The kind of extraction the waterpoint uses
extraction_type_group - The kind of extraction the waterpoint uses
extraction_type_class - The kind of extraction the waterpoint uses
management - How the waterpoint is managed
management_group - How the waterpoint is managed
payment - What the water costs
payment_type - What the water costs
water_quality - The quality of the water
quality_group - The quality of the water
quantity - The quantity of water
quantity_group - The quantity of water
source - The source of the water
source_type - The source of the water
source_class - The source of the water
waterpoint_type - The kind of waterpoint
waterpoint_type_group - The kind of waterpoint
status_group - TARGET VARIABLE. The functional state of the waterpoint

FEATURE EXPLORATORY ANALYSIS

The exploration and understanding of features is crucial to developing a good predictive model. The data in our analysis contains both numerical and categorical features. We will explore each feature to understand the data represented and identify missing values, misclassified values, and information that is repeated in multiple features. Additionally, we will do feature engineering, where we create additional features from existing ones to enrich our analysis. This exploratory analysis will aid us in selecting the features that are most relevant to build our model.

Null Values, Incorrect Values, Redundant Information

At first glance, the numeric features appear to not contain null values but further inspection reveals this is incorrect. Features such as ***amount_tsh*** has 52,049 observations encoded as 0, meaning that 70% of total waterpoints have zero total static head which would make a pump useless unless the waterpoint works by gravity. We don't get useful information from a feature like **num_private** where 98.7% of the values are zero.

Similarly, we uncovered inconsistencies in some features such as ***longitude*** where a value of 0°E is not possible for Tanzania. This feature is not included in the analysis because we can use other features such as ***basin*** and ***region_code*** to indicate location.

The features ***region_code*** appears to be a proxy for the categorical column ***region***. Similarly, ***district_code*** appears to not be hierarchically defined to ***regions***, hence one of these features is probably redundant.

In the ***population*** feature, 36% of rows have value 0, which could be an error or actual missing information. Due to the nature of the values we do have in ***population***, we can replace zeros with the median.

For ***construction_year***, 36% of entries have 0 value which represents missing information on a waterpoints build date. These values will be replaced by the the average construction year of all waterpoints so we can make better estimations.

The feature ***funder*** reveals most projects are funded by the Government of Tazmania (15.2%). Since there are over 2,140 other funders representing very small percentages and the feature has some missing values, we consider not including it in the analysis.

Similarly, for ***installer*** we found that DWE was the installer in 29.3% of the cases and we had 2,411 other installers. For ***wpt_name*** the majority of waterpoints are labeled as none (6%). We will not include these features in the analysis.

FEATURE EXPLORATORY ANALYSIS

62.45% of values in ***subvillage*** are missing. This feature will be removed considering not one subvillage accounts for more than 1% of the data.

The feature ***public_meeting*** contains some missing information which we will label as unknown, given that it is hard to assume if a public meeting took place for a specific waterpoint.

The feature ***recorded_by*** is uninformative as *GeoData Consultants Ltd* is the only value for all observations.

In ***scheme_management***, 62% of the waterpoints are managed by *VWC*. VWC and four other management companies account for 85% of the waterpoints.

About half of the features in ***scheme_name*** are *missing*. We will discard both ***scheme_name*** and ***scheme_management*** from the analysis.

We will make the assumption that the 5% of values missing in ***permit*** represent the permit is absent.

For ***extraction_type,*** 45% of the waterpoints are labeled as extraction by gravity. Other categories present in this variable represent different pump types and we also have some mislabeled items.

The features ***extraction_type_group***, ***extraction_type***, and ***extraction_type_class*** contain very similar information. We will use ***extraction_type_class*** since it is the most comprehensive feature of the three.

The feature ***management_group*** has values listed as *unknown*.

The same information is present in ***payment*** and ***payment type***. There is missing information labeled as unknown. We only need to keep one of these features, **payment**.

Other features have the similar information, as is the case with ***quality_group, water_quality*** and ***quantity_group, quantity***. Similarly, ***source***, ***source_type*** and ***source_class*** have data that is similarly classified. We will keep ***quality_group, quantity***, ***source_type*** and ***source_class***, though it may have missing information hidden in the *other* category.

The feature ***gps_height*** is categorized for waterpoints *below0* and *above0*. The missing information in this category is replaced by the average of the district the waterpoint belongs to.

FEATURE EXPLORATORY ANALYSIS

The feature ***basin*** allows us to classify the waterpoints by hydrographic source and gives us valuable information that is easy to interpret.

The features ***ward*** and ***lga*** are not useful because of the number of categories in ***ward*** and the lack of structure in ***lga***, thus are not included.

Between ***waterpoint_type*** and ***waterpoint_type_group*** we will retain ***waterpoint_type*** as it distinguishes between single and multiple standpipes.

Feature Generation

Feature generation is the process through which we create new features from existing ones to enrich our analysis.

The feature ***date_recorded*** is transformed so can extract additional features ***year_recorded***, and ***month_recorded.*** When we inspect ***year_recorded*** we can identify that most water pumps were installed between 2011 and 2013.

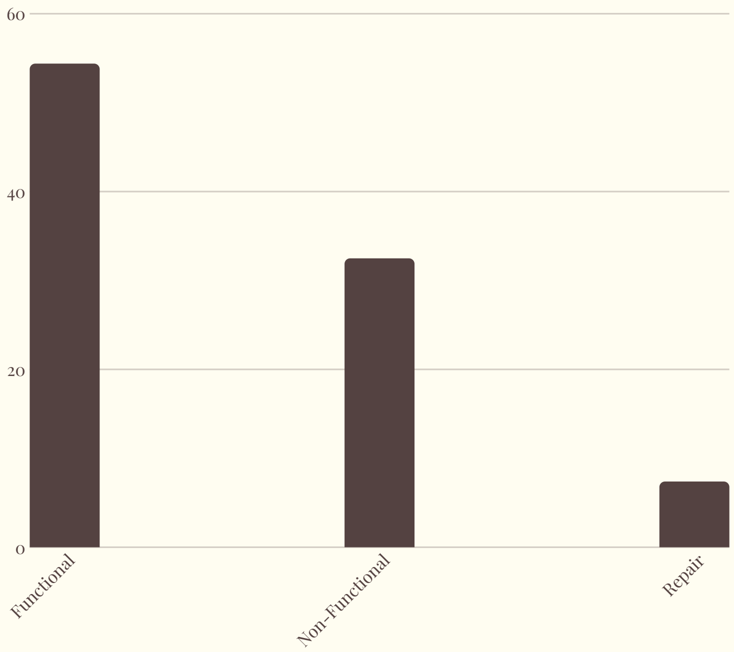
We created a variable ***season*** to represent the seasons in Tanzania: short dry (Jan-Feb), rainy (Mar-May), dry (Jun-Oct), and short rainy (Nov-Dec) based on the month recorded. We determined this under the assumption that ***date_recorded*** refers to the time when the water pump visit took place and measurements were taken. Weather and seasonal factors could influence the metrics recorded, thus will be relevant for our analysis.

The ***construction_age*** is a feature we generated from the waterpoint's ***construction_year***.

WATERPOINT STATUS

Training Set

The distribution of the target variable ***status_group*** reveals that 32,259 (54.3%) of the waterpoints are reported as functional, 22,824 (32.4%) are non-functional, and 4,317 (7.3%) need repair.



BASELINE MODEL

Baseline Model

We only used the features we believe to be informative and removed all observations with missing values to develop a basic baseline logistic regression model. This baseline model serves as a point of comparison to build upon in our machine learning selection process. Our baseline model uses **98 features** and gives us classification **accuracy of 72.4%**.

This model tells us the most important features to determine pump status are if the waterpoint is from Lake Tanganyika, month is February, payment type is other, source type is spring, region is Pwani.

Baseline Model after Feature Engineering

Once this baseline is determined we can proceed with cleaning, selecting and engineering our features as explained in the previous section. We also inspect outliers, or extreme values, present in features such as ***gps_height***, ***amount_tsh***, ***population_impute_mean***, ***population_impute_median***, ***construction_age***. Removing outliers would remove 25% of the data we need to train our model, therefore we left them in place, relying on the engineered features to perform better than the original ones. Additionally, we checked numeric features to ensure there is no correlation, which would indicate if any features directly impact or depend on each other and could result in our model producing misleading results. Our baseline logistic regression model after feature engineering uses **95 features** and gives us an **accurary of 73.1%**

After doing some clean up, the most relevant features to predict water pump functioning are region Ruvuma, season short rainy, month May June November or January, source type spring and pump static head present.

Principal Component Analysis (PCA)

While we determined earlier that the features in our analysis are not correlated, we tested out PCA to determine if by reducing the features we could improve our model's performance. We found that two principal components could explain about 99% of the variance, hence, we tested reducing from 5 numeric features to 2 principal components. However, doing this decreases our accuracy score significantly to **54.3%** so we decided not to apply PCA.

FILTERING METHODS

With filtering methods we can select only the features that are good predictors of water pump status and reduce the number of variables in our model. Two ways we can filter out the features by significance are ChiSquared Independence Test and Information Gain. In Chi-Squared selection, starting with all the features a significance threshold is determined to only keep features that are representative enough and remove the features with less importance. Information Gain, on the other hand only adds features to the analysis based on how much information the feature gives us about the class, the significant features are added based on a threshold.

Chi-Squared Selection

Using Chi-Squared selection, 24 features are removed leaving us with **72 features**, and an **accuracy score 72.7%**. According to this filtering methodology, the most important feature to determine water pump failure is a **low level of water**. The second variable selected is related to pumps with no static head, possibly **gravity class pumps**.

Information Gain Selection

Using Information Gain selection, 24 features are also removed leaving us with **72 features**, and an **accuracy score 72.8%**. According to this filtering methodology, the most important features are very similar to those achieved through the Chi-Squared Selection. Information Gain also considers the **age of the pump** a strong influencing factor to determine pump failure. If we remove the feature ***amount_tsh_bin*** we are left with **70 features** and accuracy score is **72.9%**

EMBEDDED METHODS

Ridge and Lasso

Another way to improve the performance is to fine tune the model itself. We can test out Ridge and Lasso embedded methods now that we have narrowed down the features through filtering. With this we seek to improve the accuracy of our model.

Applying Ridge to our linear regressor, our accuracy score is **72.7%**

Using Lasso, our accuracy score is **72.8%**

With Lasso, the following features: water pumps from the Rukwa and Ruvuma regions, and from Wami and Rufiji basins are considered to be important indicators of water pump status.

MODELLING AND EVALUATION

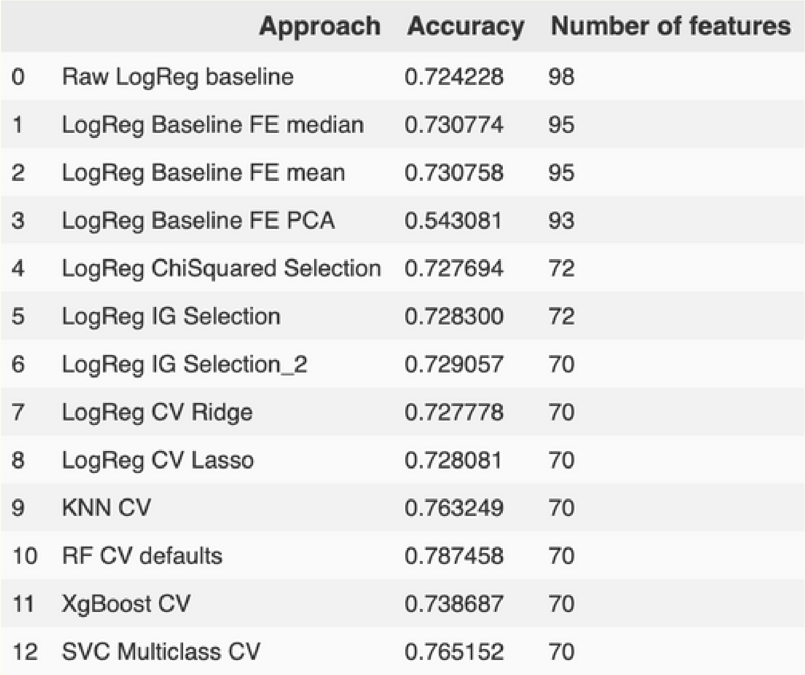
Now that we have selected the variables we will use in our analysis it is time to test different algorithms and measure their performance. Because we are seeking to classify the functioning state of the water pumps we will start with a KNN classifier, which in simple terms works under the assumption that similar things are in proximity of each other and would label a water pump's status based on how closely its characteristics are with others of its class. With KNN, we have an accuracy score of **76.3%.**

Another good family of models to predict a classification problem are ensemble models based on decision trees. Within this family we will test Random Forest and XGBoost. In a decision tree, features are used based on how important predictors they are to the status of a water pump. Then, these important features we can split the data into decision branches that help narrow down in a flow-chart fashion the classification of the water pump. Random forests apply this principle but with many decision trees at the same time. Our Random Forest classifier gave us accuracy score of **78.7%**.

XGBoost learns from the errors of initial models to improve the new predictions, with this model we achieve an accuracy score of **73.9%**.

Finally, we tested out Support Vector Classifier (SVC) which works well in classification problems by looking at the data from many different dimensions. With SVC our accuracy is **76.5%**.

The table below explains the models we tested, their accuracy score and the number of features used.

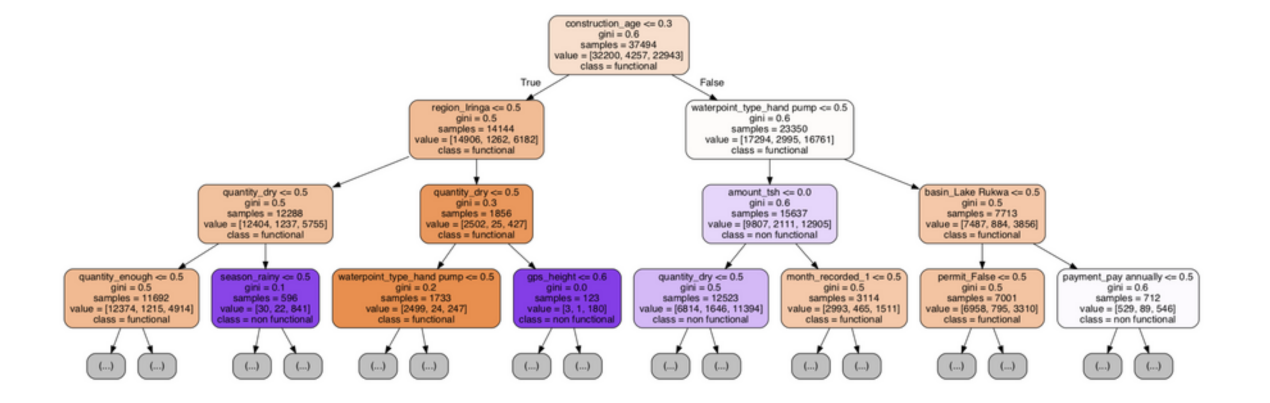


	Approach	Accuracy	Number of features
0	Raw LogReg baseline	0.724228	98
1	LogReg Baseline FE median	0.730774	95
2	LogReg Baseline FE mean	0.730758	95
3	LogReg Baseline FE PCA	0.543081	93
4	LogReg ChiSquared Selection	0.727694	72
5	LogReg IG Selection	0.728300	72
6	LogReg IG Selection_2	0.729057	70
7	LogReg CV Ridge	0.727778	70
8	LogReg CV Lasso	0.728081	70
9	KNN CV	0.763249	70
10	RF CV defaults	0.787458	70
11	XgBoost CV	0.738687	70
12	SVC Multiclass CV	0.765152	70

MODEL SELECTION

Upon inspection of the final results table, it is noteworthy that many of the evaluated models have accuracy metrics within a narrow range. Hence, the final model choice needs to accommodate an equilibrium between simplicity, computational load, economy and clarity of features, and reliability when processing new unseen data. In this sense, the **Random Forest** is favoured over all other candidates.

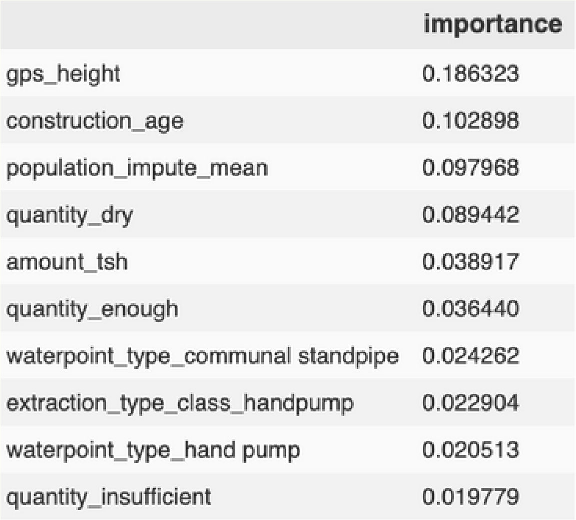
As explained earlier, Random Forest learns from many decision trees at the same time. If we inspect one tree from the forest with a depth of 5, we can visualize the features that come into play in our model.



Most Important Features

The most important features the Random Forest Model considers to predict the status of the water pump are:

- the well's altitude
- age of water pump
- population surrounding the site
- low quantity of water
- communal strandpipe waterpoints, and
- hand pump extraction



	importance
gps_height	0.186323
construction_age	0.102898
population_impute_mean	0.097968
quantity_dry	0.089442
amount_tsh	0.038917
quantity_enough	0.036440
waterpoint_type_communal standpipe	0.024262
extraction_type_class_handpump	0.022904
waterpoint_type_hand pump	0.020513
quantity_insufficient	0.019779

CONCLUSIONS

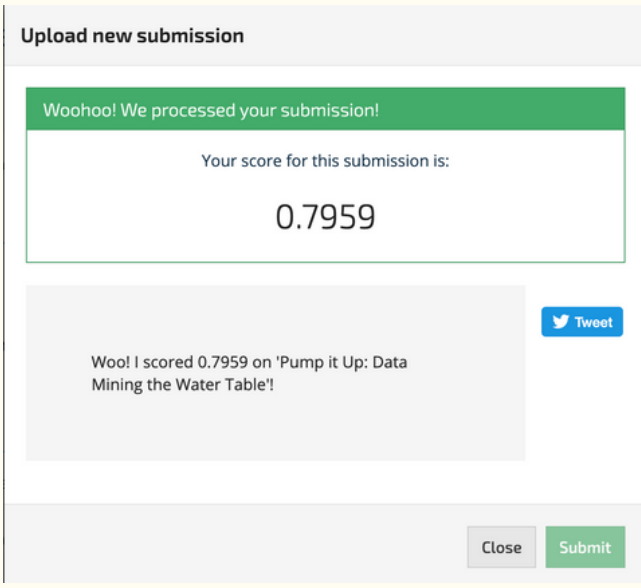
By using a Random Forest Model, we can predict the status of a water pump with 78.7% accuracy on our training data. This solution provides adequate predictive measures to a water point's status and can aid in recognizing problems in advance of a pump failure and secure the supply of potable water to all the regions in Tanzania.

Understanding the features that are most relevant predictors of water pump status for our model can help those in charge of the water pump's operation design a preventive maintenance plan to avoid failures in advance and reduce the impact a shortage of water can have on the population. For instance, assessing pumps during the dry season when water levels are low or insufficient can prevent more expensive repairs if the pump fails or breaks. It would also be relevant to monitor aging water pumps and understand that the functioning varies depending on the seasons and consider replacing them in advance instead of waiting for them to fail. It might also be worth looking closely at the functioning of units that use hand pump extraction methods. It might be feasible to install sensors in specific water points, especially those that serve larger populations, to enrich the quality of the data collected.

We encourage the Government of Tanzania to take the insights obtained in this analysis to improve the operation of water pump systems installed in the districts and regions that depend on this valuable natural resource.

Driven Data Challenge Submission

Our group submitted the prediction of our chosen model to the Driven Data Challenge under the group name ***Team G*** (IE MBD-S2 Team G, for the Machine Learning 2 assignment) and received a **score of 0.7959.**



Upload new submission

Woohoo! We processed your submission!

Your score for this submission is:

0.7959

Woo! I scored 0.7959 on 'Pump it Up: Data Mining the Water Table'!

CITATIONS

Driven Data Challenge
https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/

United Republic of Tanzania Ministry of Water
Retrieved from https://www.maji.go.tz/

Taarifa
Retrieved from http://taarifa.org/

Cover Image
Pump image courtesy of flickr user christophercjensen

Image Conclusions Page
Community waterpoint image courtesy of Tanzania Ministry of Water