SPRING 2017

PUMP IT UP

DATA MINING THE WATER TABLE

BY: ADITYA GUPTA & REKHANSH PANCHAL

COLLEGE OF COMPUTING AND INFORMATICS
UNIVERSITY OF NORTH CAROLINA, CHARLOTTE

PROJECT OVERVIEW

Dataset: Water Pump Data

Source: drivendata.org (Competition: Pump It Up)

Problem to be addressed:

Can you predict which water pumps are faulty?

- Predicting which pumps are functional, which need some repairs, and which don't work at all – Using data from Taarifa and the Tanzanian Ministry of Water.
- Predicting one of these three classes based on several variables about what kind of pump is operating, when it was installed, and how it is managed.
- A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

How?

- Following CRISP-DM process by applying concepts of Knowledge Discovery in Databases
    - Business Understanding
    - Data Understanding
    - Data Preparation
    - Modeling
    - Evaluation
    - Deployment

Business Understanding

- Tanzania is currently facing water problems.
- Only 53% of population has access to clean water.
- People rely on three major lakes and ground water delivered through water pumps.
- Find factors and variables on which functionality status depends.
- It is very difficult for people to find access to clean, sanitary water if they don't live near one of the three major lakes that border the country.
- Thus, Tanzania's groundwater is the major source of water for the nation's people; However, it is not always clean.
- Many of these groundwater wells are located near or next to toxic drainage systems, which leak into the fresh groundwater and contaminate it.
- Consequently, Tanzanians turn to surface water which contains things like bacteria or human waste; and people have no choice but to drink from, bathe in or wash their clothes in these areas.
- According to the Tanzania National Website, waterborne illnesses, such as malaria and cholera "account for over half of the diseases affecting the population," because people don't have access to potable water options.
- This project is an attempt to understand and contribute towards the water pump problem faced by Tanzania.

An image showing women taking water from a water body in Tanzania
Image Source: Humano Sphere

Data Understanding

- Dataset details

  Training set values — independent variables for the training set
  Training set labels — dependent variable (status_group) for each row.
  Test set values — independent variables that need predictions

- Variable details

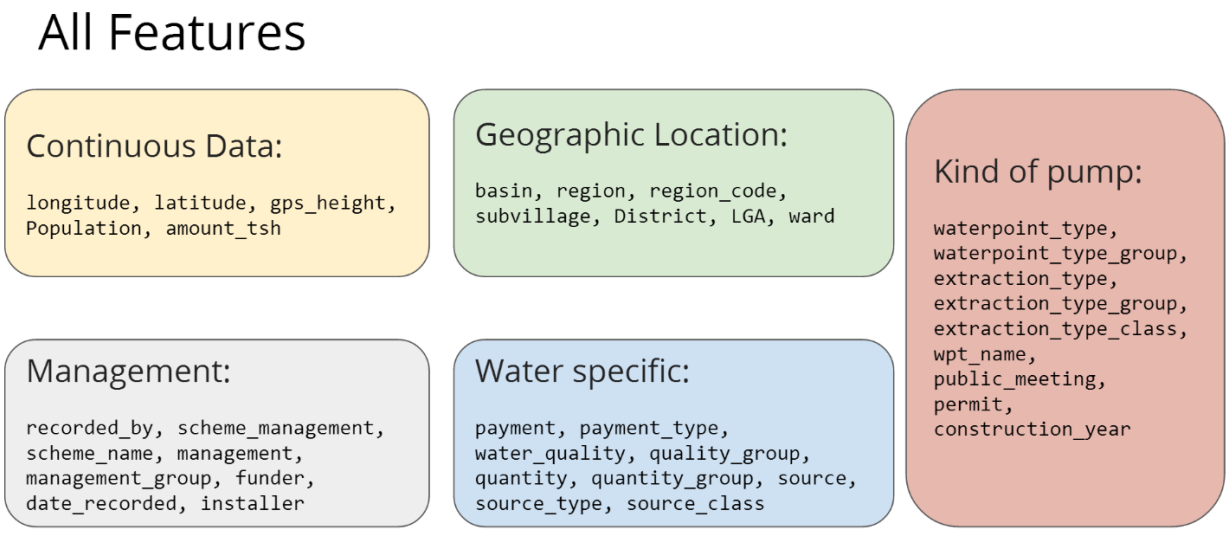
  The training dataset contains 40 variables with 59,400 data entries.

- Type of variables

category	variable name	description	Example (first row)
Water point	waterpoint_type	The kind of waterpoint	other
Water point	waterpoint_type_group	The kind of waterpoint	other
Water point	funder	Who funded the well	Dmdd
Water point	gps_height	Altitude of the well	1996
Water point	installer	Organization that installed the well	DMDD
Water point	wpt_name	Name of the waterpoint if there is one	Dinamu Secondary School
Water point	basin	Geographic water basin	Internal
Water point	population	Population around the well	321
Water point	public_meeting	True/False	TRUE
Water point	num_private	number of private	0
Water point	scheme_management	Who operates the waterpoint	Parastatal
Water point	scheme_name	Who operates the waterpoint	
Water point	permit	If the waterpoint is permitted	TRUE
Water point	construction_year	Year the waterpoint was constructed	2012
Water point	extraction_type	The kind of extraction the waterpoint uses	other
Water point	extraction_type_group	The kind of extraction the waterpoint uses	other
Water point	extraction_type_class	The kind of extraction the waterpoint uses	other
Water point	management	How the waterpoint is managed	parastatal
Water point	management_group	How the waterpoint is managed	parastatal

category	variable name	description	Example (first row)
Water point	longitude	GPS coordinate	35.2907992
Water point	latitude	GPS coordinate	-4.05969643
Water point	subvillage	Geographic location	Magoma
Water point	region	Geographic location	Manyara
Water point	region_code	Geographic location (coded)	21
Water point	district_code	district_code - Geographic location (coded)	3
Water point	lga	Geographic location	Mbulu
Water point	ward	Geographic location	Bashay
Water	amount_tsh	amount water available to waterpoint	0
Water	payment	What the water costs	never pay
Water	payment_type	What the water costs	never pay
Water	water_quality	The quality of the water	soft
Water	quality_group	The quality of the water	good
Water	quantity	The quantity of water	seasonal
Water	quantity_group	The quantity of water	seasonal
Water	source	The source of the water	rainwater harvesting
Water	source_type	The source of the water	rainwater harvesting
Water	source_class	The source of the water	surface
Record	id	index of the observation	50785
Record	date_recorded	date_recorded - The date the row was entered	2/4/13
Record	recorded_by	recorded_by - Group entering this row of data	GeoData Consultants Ltd

Overview of all features:



Data Preparation

- Merging Dataset (value and labels)
    - The two training datasets were merged using ID as the key.

- Exploratory Data Analysis
    - Streamlining the variables.
        - During EDA, we noticed 7 pairs of independent categorical variables consist of identical or very similar levels to each other and 9 variables turned out to be location based.
        - We evaluated the relationship among the variables to avoid duplicate information and separate the variables that are distinct and meaningful.

    - Interrelated Variables
        - water_quality & quality_group
        - extraction_type & extraction_type_class & extraction_type_group
        - quantity & quantity_group
        - management & management_group
        - payment & payment_type
        - source & source_type & source_class
        - waterpoint & waterpoint_type_group

    - Location Variables
        - Basin
        - Region
        - region_code
        - district_code
        - lga
        - longitude
        - latitude
        - ward
        - subvillage

    - Detecting outliers
        - According to exploratory data analysis, the five variables contained outliers. Most of values within those columns were concentrated at ‘zero’ from density plot.
        - As 98.7% of num_private are filled with zero, we conclude that this variable is hard to be imputed accurately.
        - We decided to try to impute the amount_tsh and population although 70% of amount_tsh are zero and 47% of population are zero or one since these proportions are still imputable and there was no significant pattern in outliers when they are applied to the target variable.

Variable	Number of the value Zero	% in Total	P-value from chi-square test
num_private	58643	98.7%	Below 0.01
amount_tsh	41639	70%	Below 0.01
population	28406*	47.8%*	Below 0.01
construction_year	20709	34.9%	Below 0.01

- Variables with missing values
  - funder
  - installer
  - subvillage
  - public_meeting
  - scheme_management
  - scheme_name
  - permit
- Chi-Square test was performed before imputing them test for any significant pattern and decide whether to impute or not.
- According to the test none of the variables with missing values or unknown
- Values have any meaningful pattern and resulted in below 0.01 p-values.

variable	number of null	% of null
funder	3635	6.1%
installer	3655	6.2%
subvillage	371	0.6%
public_meeting	3334	5.6%
scheme_management	3877	6.5%
scheme_name	28166	47.4%
permit	3056	5.1%

Number of *null* values in variables.

variable	number of 'unknown'	% of 'unknown'
quantity	789	1.3%
quantity_group	789	1.3%
management	561	0.9%
management_group	561	0.9%
payment	8157	13.7%
payment_type	8157	13.7%
source	66	0.1%
source_class	278	0.5%

Number of *unknown* values in variables.

Variables with unknown values

- quantity
- quantity_group
- management
- management_group
- payment
- payment_type
- source
- source_class
- quality_group
- water_quality

Variable	Number of the value Zero	% in Total	P-value from chi-square test
num_private	58643	98.7%	Below 0.01
amount_tsh	41639	70%	Below 0.01
population	28406*	47.8%*	Below 0.01
construction_year	20709	34.9%	Below 0.01

Chi-square test result

- Missing values exist as null value or 'unknown' in 15 variables accounting for up to 47% of observations.
- Among 6 continuous variables, population, num_private, and amount_tsh contain outliers accounting for more than 7% of every observation
- Two continuous variables, amount_tsh and num_private, are the most skewed and deviated from each other

Skewness and Kurtoisis

	gps_height	population	latitude	longitude	amount_tsh	num_private
Skewness	0.462	12.660	-0.152	-4.191	**57.806**	**91.931**
Kurtosis	-1.292	402.246	-1.058	19.185	**4903.130**	**11136.357**

- Skewness
  - For normally distributed data, the skewness should be about 0.
  - A skewness value > 0 means that there is more weight in the left tail of the distribution

- Kurtosis
  - Kurtosis is the fourth central moment divided by the square of the variance. We used the Fisher's definition (normal ==> 0.0).
  - Therefore, if all values are equal, return -3 for Fisher's.

- Detect class-imbalance variable
  - Among the categorical independent variables after streamlining ten of them have one dominant level which accounts for more than 25% of the total observations.
  - To examine the impact of such imbalance in class on the target variable, we conducted the chi-square test. We divided each variable into two groups - one with major level and another with the rest of levels.
  - Thus, such skewness is proved to have no impact on the target variable and we decided to leave them as they are.

Chi Square Test

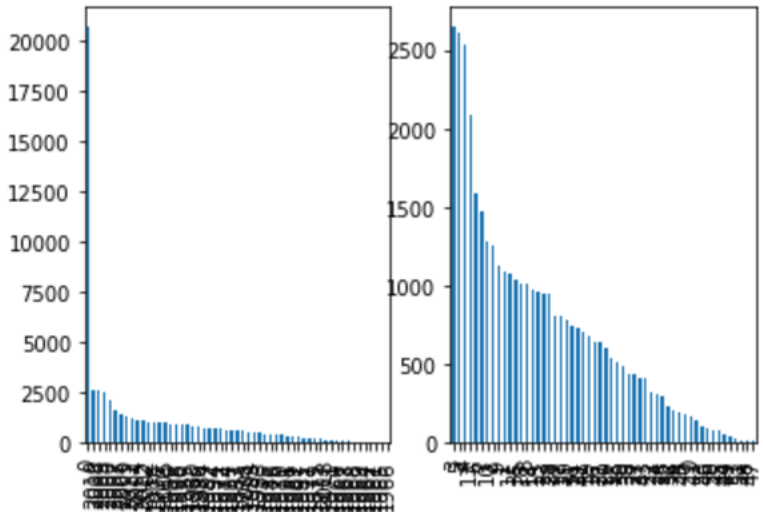
Variable	Number of Major level	% of Major level in Total	P-value chi-square test
scheme_management	VWC	61.9%	Below 0.01
quality_group	good	85.6%	Below 0.01
quantity	enough	55.9%	Below 0.01
waterpoint_type	Communal standpipe	48%	Below 0.01
installer	DWE	29.2%	Below 0.01
extraction_type_class	gravity	45%	Below 0.01
management_group	user-group	88.4%	Below 0.01
payment	never pay	42.7%	Below 0.01
permit	True	65.4%	Below 0.01
source_class	groundwater	77%	Below 0.01

Data Munging

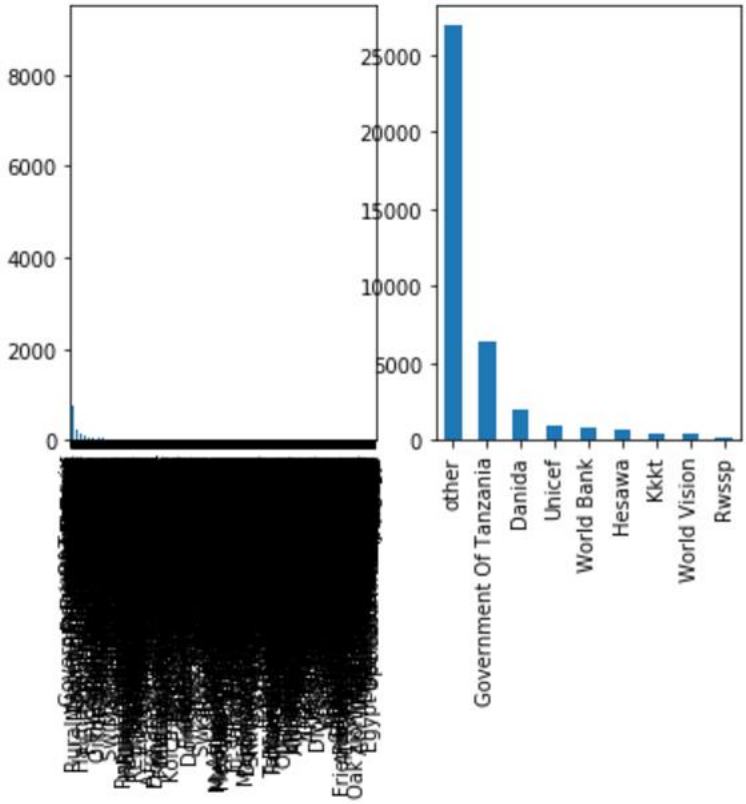
- We dropped funder, installer, and scheme_name categorical variables as they had more than 100 levels.
- We found that few columns contain null values. Hence, we started fixing them before beginning any sort of Analysis on the data.
- Starting with value “funder”
  - We categorized the variable by replacing less frequent values by “*other*”.
- Exploring "installer" variable
  - We categorized the variable by replacing less frequent values by “*other*”.
- Subvillage
  - As there are so many unique values, high frequency values will not be able to dominate other variables.
  - Hence, we removed the column from dataset
- public_meeting
  - It has missing values. Hence, we have filled those values with “*Unknown*”.
- scheme_management
  - Categorizing the variable by replacing less frequency values with “*other*”.
- recorded_by
  - Removing this column as it has only one level.
- amount_tsh
  - It has 41639 entries filled with zero, thus dropping this column.
- permit
  - Fill missing values using NA
- scheme_name
  - We have removed scheme_name as there are too many different scheme types.

We have dropped the following variables as they were redundant entries:

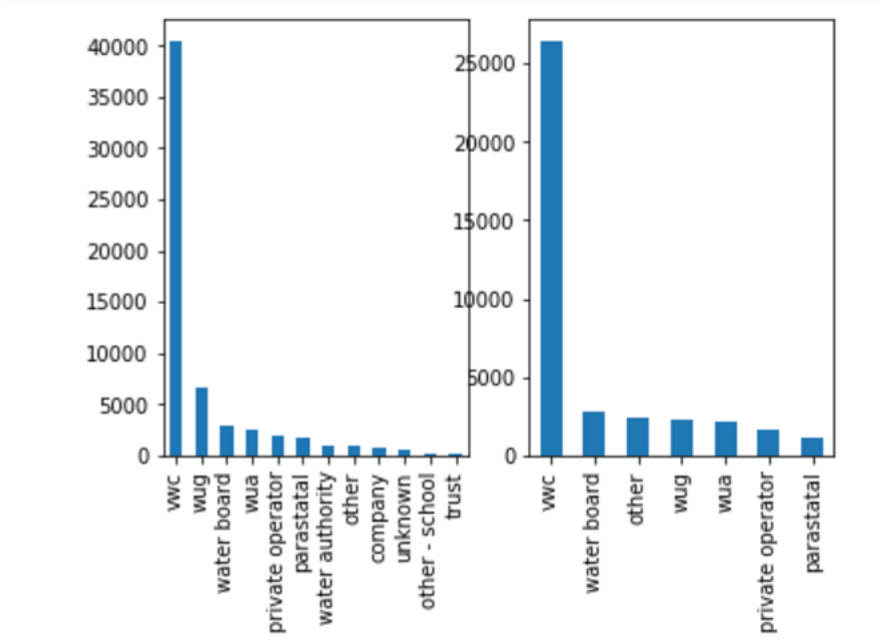
- payment
- quality_group
- quantity_group
- waterpoint_type_group
- source
- eource_class
- extraction_type
- extraction_type_group
- management_group
- num_private
- lga, longitude, latitude. region_code
- population



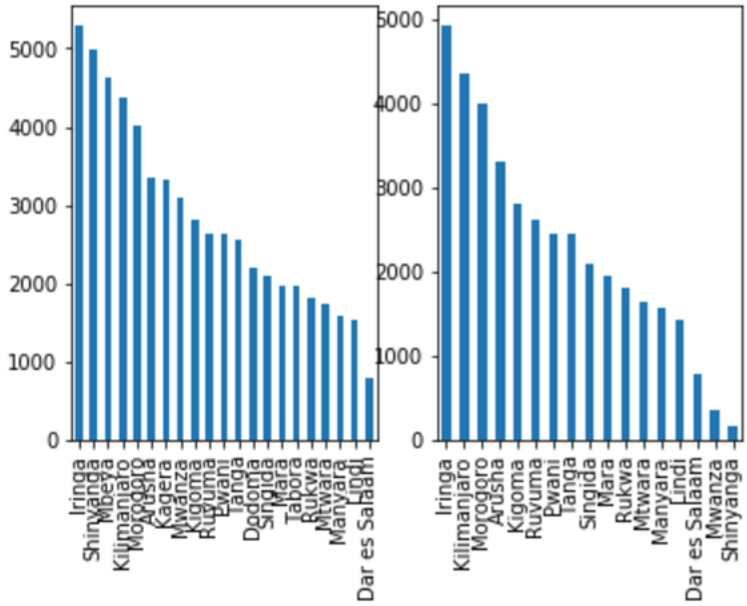
Construction year



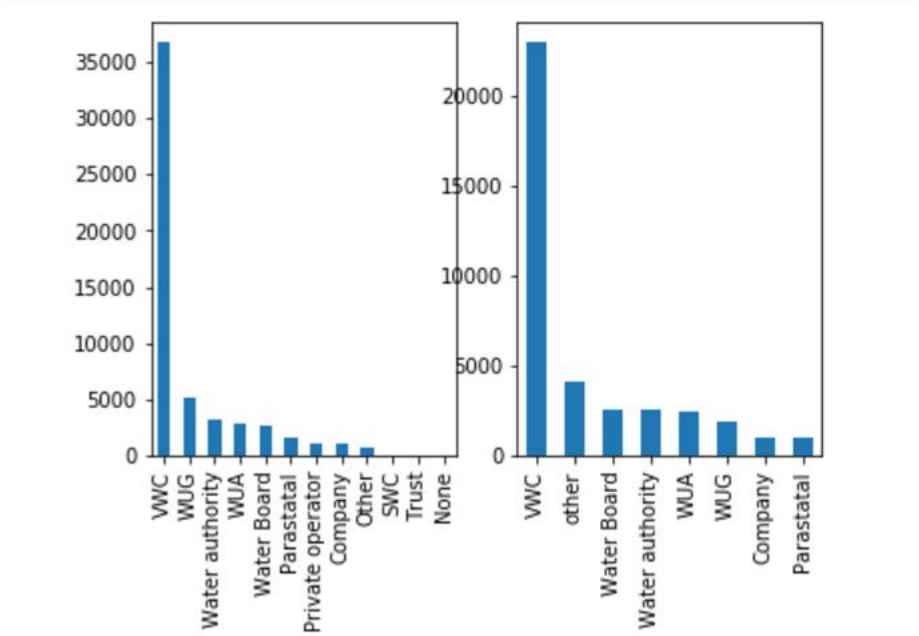
Funder



management



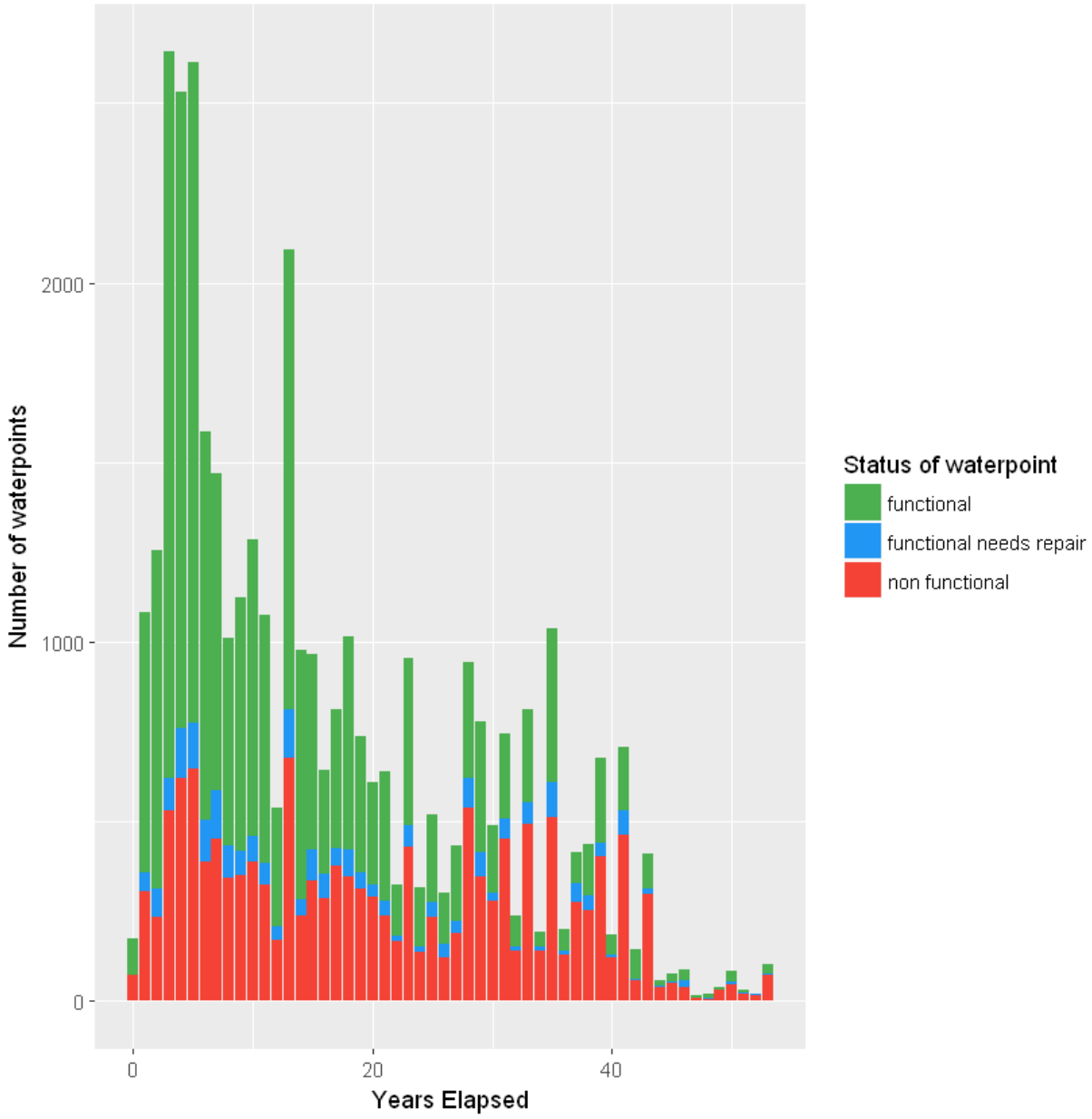
region



scheme management

We saved the resulting data frame in to a new csv file named ***cleanedTraining.csv***

Before moving to Modeling phase, let's have a look at status of pumps in training data set.



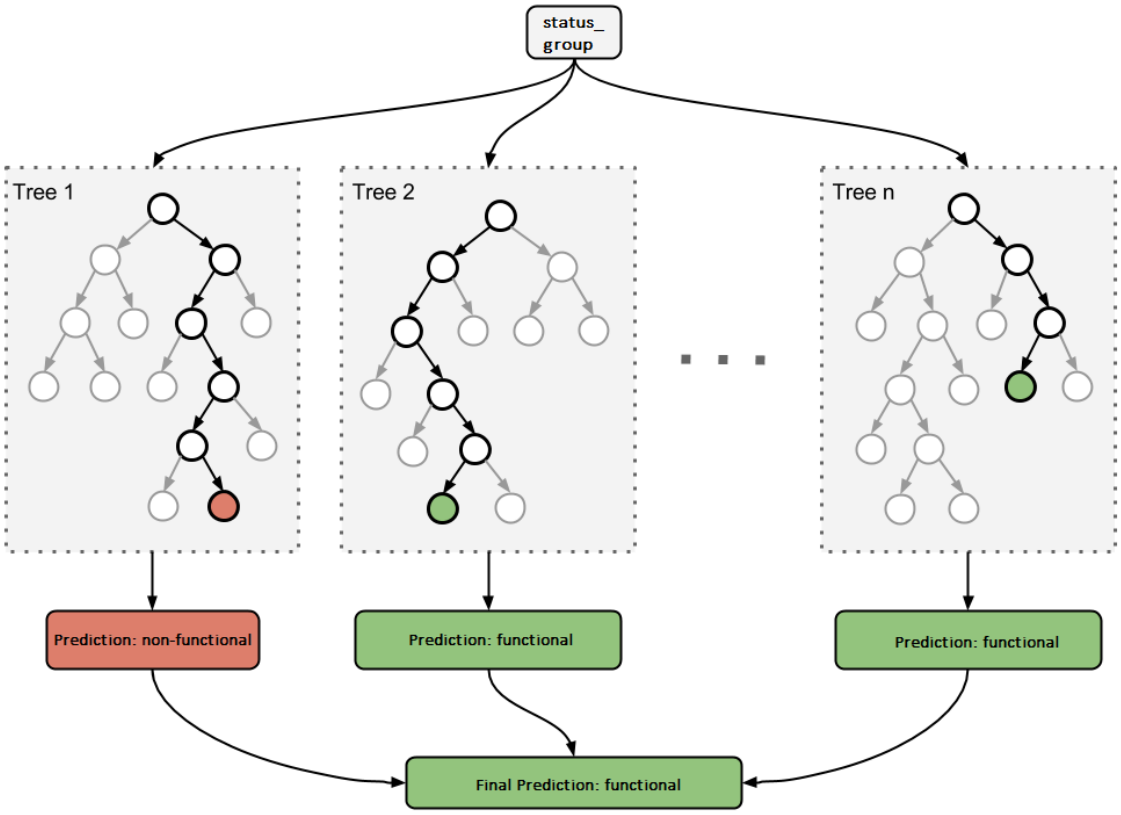
We can visualize the insight the recently built water pumps are more likely to be functional.

Data Modeling

- In this phase, used two modeling techniques, namely Random Forest Classification and Multinomial Logistic Regression.
- Both the techniques were employed to find the parameters which are important for classification of functional status of water pumps
- Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.
- Random Forest Classification was found to have better prospects will classify target variable. Hence, we considered this technique over another one.
- Multiple iterations of Random Forest were performed over different sizes of trees and nodes.
- Tree size from 1 to 1000 and node size 2 to 35, were iterated to find a best fit model with parameters as mentioned below:
    - Tree Size = 100
    - Node Size = 2
    - Predictor Variables = 15

Predictor Variables
Funder
Installer
Basin
Region
Public Meeting
Permit
Construction Year
Extraction
Management
Payment Type
Water Quality
Quantity
Source Type
Water Point Type

Random Forest Classification Overview



Random Forests:

- The random forest is an ensemble approach that can be thought of as a form of nearest neighbor predictor.
- Ensembles are a divide-and-conquer approach used to improve performance.
- The main principle behind ensemble methods is that a group of “weak learners” can come together to form a “strong learner”. Each classifier, individually, is a “weak learner,” while all the classifiers taken together are a “strong learner”.

Trees and Forests

- The random forest starts with a standard machine learning technique called a “decision tree” which, in ensemble terms, corresponds to our weak learner. In a decision tree, an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets.
- For details see here, from which the figure above is taken.

- The tree advises us, based upon various parameters and conditions, the functional status of water pump. For example, if the water quality is bad and the management is bad and quantity is dry, then it's probably non-functional pump.
- The random forest (as seen above) takes this notion to the next level by combining trees with the notion of an ensemble. Thus, in ensemble terms, the trees are weak learners and the random forest is a strong learner.
- The predictor variable that provides the best split, according to some objective function, is used to do a binary split on that node.

Running a Random Forest

- When a new input is entered in the system, it is run down all the trees. The result may either be an average or weighted average of all the terminal nodes that are reached, or, in the case of categorical variables, a voting majority.

Creating a random Forest model

```
model_forest <- randomForest(status_group ~
                             funder +
                             installer +
                             basin +
                             region +
                             quantity +
                             public_meeting +
                             scheme_management+
                             permit+
                             construction_year+
                             extraction_type_class+
                             management+
                             payment_type+
                             water_quality+
                             quantity+
                             source_type+
                             waterpoint_type,
                             data = training,
                             importance = TRUE,
                             ntree = 100,
                             nodesize = 2)
```

Passing test data to the random forest model

```
# Predict the values in training using the random forest model
pred_forest_training <- predict(model_forest, test_data)
```

Importance of Predictor Variables:

```
# Evaluating the importance of each variable
importance(model_forest)
varImpPlot(model_forest)
```

	functional	functional needs repair	non functional	MeanDecreaseAccuracy	MeanDecreaseGini
funder	30.37415	17.276841	26.56106	35.07761	502.5393
installer	27.33392	13.426249	27.30042	34.96236	412.7644
basin	35.11195	17.892794	28.81280	40.91184	799.9515
region	48.24266	23.408693	44.30726	57.81436	1124.6863
quantity	86.64213	25.097695	87.32212	108.80453	2465.3448
public_meeting	22.34347	6.215242	26.61116	32.49025	237.9744
scheme_management	26.03963	14.265555	34.13837	37.48153	567.6829
permit	30.09212	10.006816	28.92950	41.47227	237.6629
construction_year	42.78693	25.598953	61.63912	59.41509	1898.0358
extraction_type_class	47.84189	19.486544	28.99462	53.62523	1077.9896
management	25.98888	13.322359	21.09754	29.90309	558.8647
payment_type	46.42553	22.225768	52.76153	61.36105	1211.6198
water_quality	19.16223	11.542349	16.38344	22.00588	322.2270
source_type	28.84730	22.525324	36.68544	37.82825	745.5785
waterpoint_type	46.39910	31.639882	36.98297	56.80179	1396.1210

Note that:

With many predictors, the eligible predictor set will be quite different from node to node.

The greater the inter-tree correlation, the greater the random forest error rate, so one pressure on the model is to have the trees as uncorrelated as possible.

Strengths and weaknesses

- Random forest runtimes are quite fast, and they can deal with unbalanced and missing data.
- Random Forest weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.
- Of course, the best test of any algorithm is how well it works upon considering data set.

Evaluation Phase

- At this stage in the project a model is built that appears to have high quality, from a data analysis perspective.
- Before proceeding to final deployment of the model, it is important to evaluate the model more thoroughly, and review the steps executed to construct the model, to be certain it properly achieves the business objectives.
- Hence, we go back to business objectives and check whether each thing is accounted for.

Finding Accuracy of the model built: **86%**

```
# Creating a confusion matrix to evaluate the model (compare the predicted labels to the actual labels)
cm <- table(pred_forest_training, test_data$status_group)
cm

# Calculate accuracy
round(sum(diag(cm))/sum(cm), 4)
```

```
pred_forest_training      functional functional needs repair non functional
  functional                    4135                     286            467
  functional needs repair         28                     139             15
  non functional                 173                      62           2342
```

0.8652

```
# Can evaluate the model with more statistics calculated automatically
library(caret)
confusionMatrix(pred_forest_training, test_data$status_group)
```

```
Warning message:
"package 'caret' was built under R version 3.3.3"Loading required package: lattice
Confusion Matrix and Statistics

                          Reference
Prediction                 functional functional needs repair non functional
  functional                     4135                     286            467
  functional needs repair          28                     139             15
  non functional                  173                      62           2342

Overall Statistics

               Accuracy : 0.8652
                 95% CI : (0.8573, 0.8728)
    No Information Rate : 0.567
    P-Value [Acc > NIR] : < 2.2e-16
```

Model Analysis with Statistics

- As seen from modeling, we have successfully classified the water pumps based on predictor variables into *functional*, *non-functional*, and *need maintenance* with an accuracy of **86%**.
- As seen from the confusion matrix, the number of **True Positives >> False Positives**, we can say that model is appropriate to be deployed.
- At the end of this phase, a decision on the use of the data mining results is reached.

Successful prediction of water pump status_group