Pump it Up: Data Mining the Water Table

November 23, 2016

Domain Background

Across Africa, cholera, typhoid, dysentery and other diseases kill thousands each year. To help the people of Tanzania(2007), The Tanzanian government, with support from UN Development Programme(UNDP), responded to the water problems by install Drinking Water Taps and Decentralised the maintenance for quick response. Today this water infrastructure is facing repair and maintenance issues causing a disconnection for drinking water needs.

This is also an intermediate-level practice competition by DrivenData.

Problem Statement

Using data from Taarifa and the Tanzanian Ministry of Water, predicting which pumps are functional, which need some repairs, and which don't work at all. Predict one of these three classes based on a number of variables about what kind of pump is operating, when it was installed, and how it is managed.

Datasets and Inputs

Dataset is collected from Taarifa and the Tanzanian Ministry of Water which are available here. Taarifa is an open source platform for the crowd sourced reporting and triaging of infrastructure related issues and Tanzanian Ministry of Water is the central governing body to Tanzania.

Dataset Features:

- Col ID: A simple index for counting columns
- UniqCount: Total number of unique labels for that feature.
- UniqVal: If UniqCount is < 30, then we shall have that columns's respective values & counts

Col ID	Col Name	UniqCount	Col Values	UniqValCount
1	amount_tsh	98		
2	date_recorded	356		
3	funder	1897		
4	gps_height	2428		
5	installer	2145		
6	longitude	57516		
7	latitude	57517		
8	wpt_name	37400		
9	num_private	65		
10	basin	9	Wami / Ruvu	5987
			Ruvuma / Southern Coast	4493
			Rufiji	7976
			Internal	7785
			Lake Rukwa	2454
			Lake Nyasa	5085
			Pangani	8940

			Lake Victoria	10248
			Lake Tanganyika	6432
11	subvillage	19287		
12	region	21	Mwanza	3102
			Kagera	3316
			Dodoma	2201
			Lindi	1546
			Morogoro	4006
			Arusha	3350
			Tabora	1959
			Iringa	5294
			Shinyanga	4982
			Dar es Salaam	805
			Mbeya	4639
			Kilimanjaro	4379
			Mtwara	1730
			Kigoma	2816

			Rukwa	1808
			Ruvuma	2640
			Manyara	1583
			Pwani	2635
			Singida	2093
			Tanga	2547
			Mara	1969
13	region_code	27	1	2201
			2	3024
			3	4379
			4	2513
			5	4040
			6	1609
			7	805
			8	300
			9	390
			10	2640

			11	5300
			12	4639
			13	2093
			14	1979
			15	1808
			16	2816
			17	5011
			18	3324
			19	3047
			20	1969
			21	1583
			24	326
			40	1
			60	1025
			80	1238
			90	917
			99	423

14	district_code	20	0	23
			1	12203
			2	11173
			3	9998
			4	8999
			5	4356
			6	4074
			7	3343
			8	1043
			43	505
			13	391
			80	12
			67	6
			53	745
			23	293
			62	109
			33	874

			60	63
			30	995
			63	195
15	lga	125		
16	ward	2092		
17	population	1049		
18	public_meeting	2	False	5055
			True	51011
19	recorded_by	1	GeoData Consultants Ltd	59400
20	scheme_management	12	None	1
			Private operator	1063
			Water authority	3153
			Water Board	2748
			SWC	97
			Parastatal	1680
			WUA	2883
			Other	766

			WUG	5206
			Trust	72
			Company	1061
			VWC	36793
21	scheme_name	2696		
22	permit	2	False	17492
			True	38852
23	construction_year	55		
24	extraction_type	18	windmill	117
			submersible	4764
			other - mkulima/shinyanga	2
			mono	2865
			climax	32
			india mark ii	2400
			afridev	1770
			gravity	26780
			walimi	48

			cemo	90
			nira/tanira	8154
			other - play pump	85
			other	6430
			other - swn 81	229
			swn 80	3670
			ksb	1415
			other - rope pump	451
			india mark iii	98
25	extraction_type_group	13	submersible	6179
			rope pump	451
			wind-powered	117
			other motorpump	122
			mono	2865
			india mark ii	2400
			afridev	1770
			gravity	26780

			other handpump	364
			nira/tanira	8154
			other	6430
			india mark iii	98
			swn 80	3670
26	extraction_type_class	7	submersible	6179
			handpump	16456
			rope pump	451
			motorpump	2987
			gravity	26780
			other	6430
			wind-powered	117
27	management	12	private operator	1971
			water authority	904
			unknown	561
			water board	2933
			parastatal	1768

			wua	2535
			other	844
			wug	6515
			trust	78
			company	685
			other - school	99
			vwc	40507
28	management_group	5	parastatal	1768
			other	943
			commercial	3638
			user-group	52490
			unknown	561
29	payment	7	unknown	8157
			never pay	25348
			pay monthly	8300
			other	1054
			pay when scheme fails	3914

			pay annually	3642
			pay per bucket	8985
30	payment_type	7	on failure	3914
			per bucket	8985
			monthly	8300
			unknown	8157
			annually	3642
			never pay	25348
			other	1054
31	water_quality	8	fluoride	200
			unknown	1876
			salty abandoned	339
			coloured	490
			fluoride abandoned	17
			salty	4856
			milky	804
			soft	50818

32	quality_group	6	good	50818
			colored	490
			unknown	1876
			salty	5195
			milky	804
			fluoride	217
33	quantity	5	dry	6246
			insufficient	15129
			enough	33186
			seasonal	4050
			unknown	789
34	quantity_group	5	dry	6246
			insufficient	15129
			enough	33186
			seasonal	4050
			unknown	789
35	source	10	unknown	66

			spring	17021
			machine dbh	11075
			lake	765
			shallow well	16824
			other	212
			rainwater harvesting	2295
			dam	656
			river	9612
			hand dtw	874
36	source_type	7	river/lake	10377
			spring	17021
			shallow well	16824
			other	278
			rainwater harvesting	2295
			dam	656
			borehole	11949
37	source_class	3	unknown	278

			groundwater	45794
			surface	13328
38	waterpoint_type	7	hand pump	17488
			communal standpipe	28522
			improved spring	784
			other	6380
			communal standpipe multiple	6103
			dam	7
			cattle trough	116
39	waterpoint_type_group	6	hand pump	17488
			communal standpipe	34625
			improved spring	784
			other	6380
			dam	7
			cattle trough	116

(ALL 39 columns's unique values counts)(98, 356, 1897, 2428, 2145, 57516, 57517, 37400, 65, 9, 19287, 21, 27, 20, 125, 2092, 1049, 2, 1, 12, 2696, 2, 55, 18, 13, 7, 12, 5, 7, 7, 8, 6, 5, 5, 10, 7, 3, 7, 6)
(542989639101365927794152062178992491838404172880830791680000000000000

(69 digits) is product of these 39 unique values, which is exponentially greater than 59K records we have.)

Input labels data has 39 Features with 59,400 rows. Although we seem to have a good data set, looking at the unique values counts from below 39 columns we can say that we could potentially encounter Curse of Dimensionality. But, as we can see some of columns pairs (extraction_type, extraction_type_group), (quantity & quantity_group), (source, source_class) seems have closer relation and column by 'recorded_by' has only one unique value. So, we might have a chance to escape Curse of Dimensionality.

Description of the Labels

The labels in this dataset are simple. There are three possible values:

- functional - the water point is operational and there are no repairs needed
- functional needs repair - the water point is operational, but needs repairs
- non functional - the water point is not operational

Col ID	Col Name	UniqCount	Col Values	UniqValCount
40	status_group	3	functional needs repair	4317
			functional	32259
			non functional	22824

Solution Statement

A smart understanding of which water points will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

We will use familiar (inherently) multi-class Supervised Classifiers like Tree Algorithms(RF/GBT)/Support Vector Machines. These are easy to train and self learning & evaluation nature make them a general good technique. During model selection we will also explore One-vs-Rest Sklearn's MultiClassification Technique. As the data is unbalanced, we believe having a One-vs-Rest might not perform well.

Benchmark Model

With a simplistic data transformation and with the help of Random Forest Classifiers, we have created a benchmark submission of 0.7970 for which source code is here

Evaluation Metrics

Accuracy Score:

As the evaluation metric of the competition use Accuracy Score /Classification Rate, we can use this metric.

The classification rate, which calculates the percentage of rows where the predicted class in the submission matches the actual class in the test set. The maximum is 1 and the minimum is 0. The goal is to maximise the classification rate.

Classification Rate = (1/N)* ∑ i=0 N I (Prediction == Actual)
Sample Example from Python Scikit

```
from sklearn.metrics import accuracy_score
y_pred = [0, 2, 1, 3]
y_true = [0, 1, 2, 3]

# calculating score
accuracy_score(y_true, y_pred)
```

Weighted Accuracy Score:

Approaching to a realistic perspective, this is calculation is to help Governing/Supporting bodies to identify the water pumps that `requires repairs` or `non functional`. So, I feel prioritising this logic identifying (being biased) these two categories is more important.

To drill down further, I would say prioritise the `functional needs repairs` over `non functional`. Thus, with in limited resources(time & money) we can solve more problems.

```
import numpy as np
from sklearn.metrics import accuracy_score
y_true = [0, 1, 0, 2]
my_sample_weight = np.array([1,2,1,5])

# Alogrithm good at predicting 0's
pred1 = [0, 2, 0, 1]

# Alogrithm good at predicting 1's
pred2 = [0, 1, 2, 1]

# Alogrithm good at predicting 2's
pred3 = [1, 1, 1, 2]

for i, pred in enumerate([pred1, pred2, pred3]):
    print 'Case %s: %f' % (i, accuracy_score(y_true, pred,
sample_weight=my_sample_weight))
```

Results:

```
Case 0: 0.222222
Case 1: 0.333333
Case 2: 0.777778
```

As many tree algorithms(self correcting nature) can use weighted samples to prioritise, we can use this logic in model training level itself and might genenrate better results during model evaluation stage.

Weighted F1 Score:

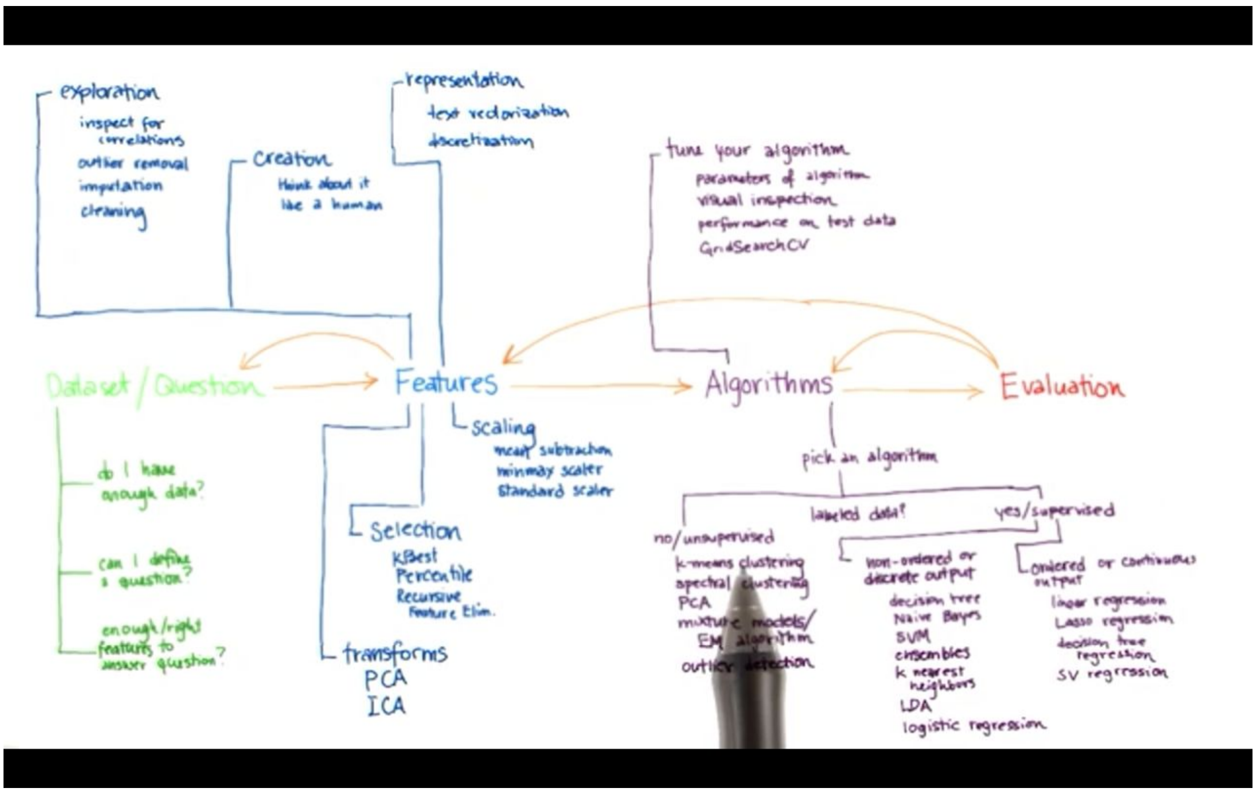
Accuracy based on the class labels(unbalanced) distribution. F1 Score will help to calculate score in proportional to data & labels.

```
from sklearn.metrics import f1_score
y_pred = [0, 2, 1, 3]
y_true = [0, 1, 2, 3]

# calculating score
f1_score(y_true, y_pred, average='weighted')
```

Project Design

As shown in below image, we are going to do a step by step development progress on here.

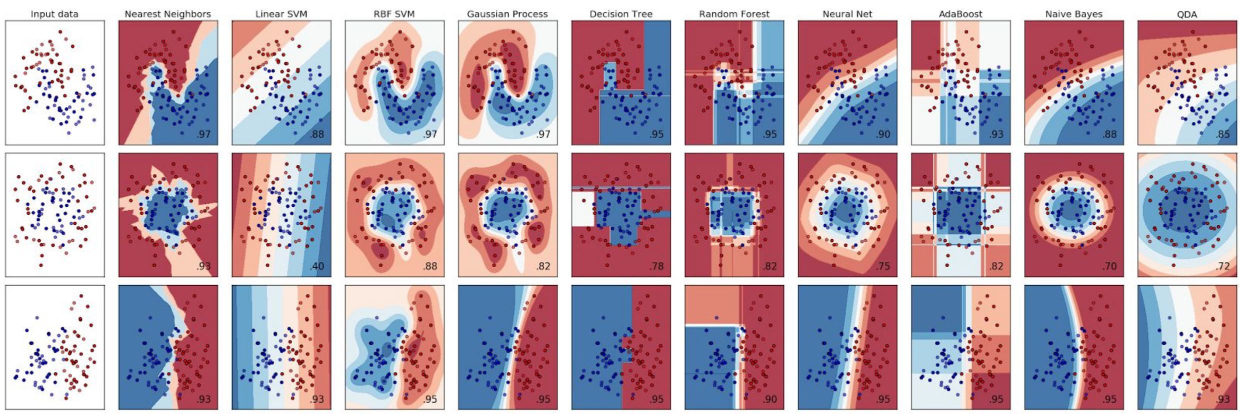


With Random Forest Classifier, we were able to generate a benchmark of 0.7970. So, first we will start with going to deeper understanding of Random Forest worked and what features contributed it to generate this score in training.

(Implementation Plan)

1. Questions on data
2. Feature Exploration

    - PCA Transformation Checking
    - Select K Best Checking
    - Exploration - outliers check
3. Algorithm Selection
    - Unsupervised Learning Exploration(Gaussian Process, Neural Nets)
    - Supervised Learning(GBT Trees, Nearest Neighbours, RF, One-vs-One)
    - Parameter Tuning
4. Evaluation. Back to 1 with.
5. Re-Evaluation with threshold improvisation check.
6. Submission



As we can see from above analysis, I find that `Nearest Neighbour` performs better when Random Forest is performing low. Also for different learning process from that of Random Forest. GBT Tree, sometime have seems performed better than Random Forest.

We will be using Gaussian Process, Neural Nets for unsupervised Learning exploration. No specific reason but taken, two models different kinds of models for exploration.

Sources & References

- DataDriven
- Submission Code
- Wikipedia: Water Supply & Sanitation in Tanzania
- UN Report
- UN 2007 Water Taps Installation
- GBT Video Lecture
- GBT
- Classifier Comparison

- Multi-class Classification
- Multi-class and multi label algorithms
- Multi-class Metric
- Standford UnSupervised Learning