Pump it Up: Data Mining the Water Table

November 14, 2016

Domain Background

Across Africa, cholera, typhoid, dysentery and other diseases kill thousands each year. To help the people of Tanzanian(in 2007), The Tanzanian government, with support from UN Development Programme(UNDP), responded to the water problems by install Drinking Water Taps and Decentralized the maintenance for quick repsonses.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:

- 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

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.

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

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

Benchmark Model

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

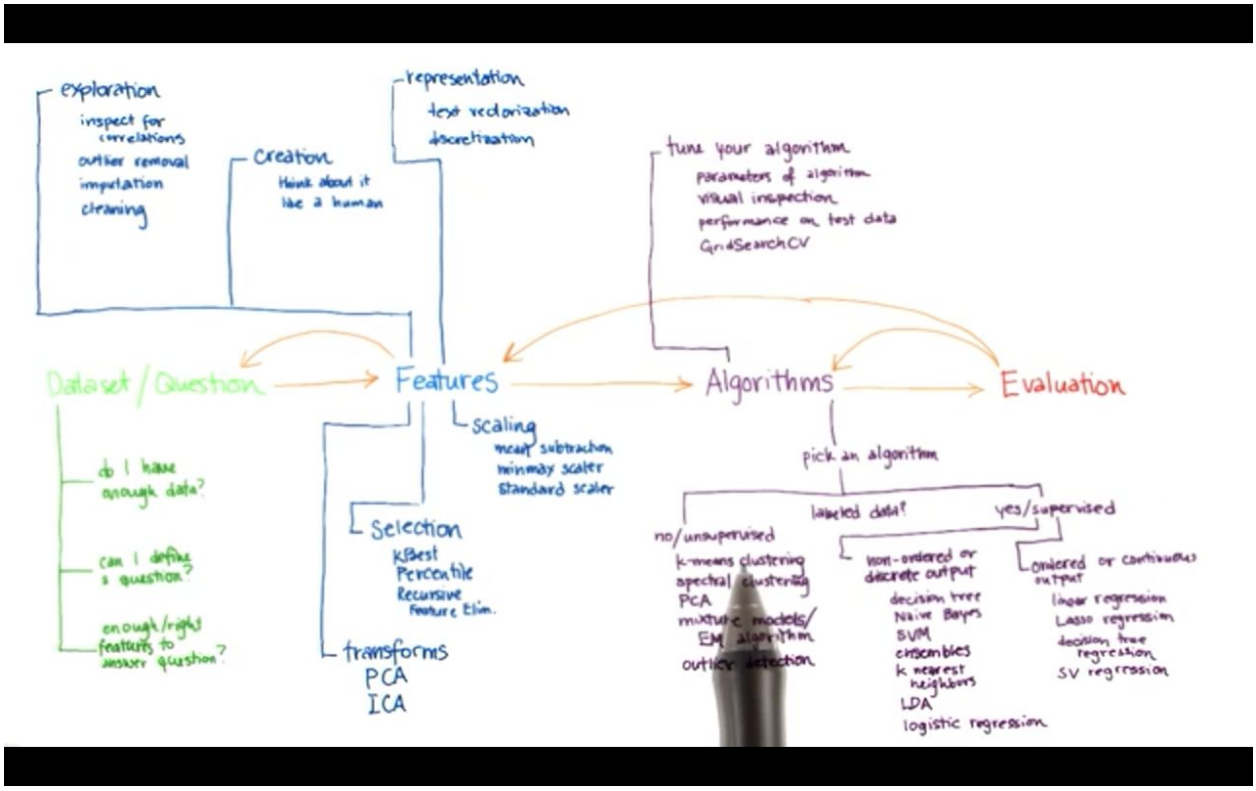
Evaluation Metrics

The metric used is 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 maximize the classification rate.

Classification Rate = (1/N)* ∑ i=0 N I (Prediction == Actual)

Project Design

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



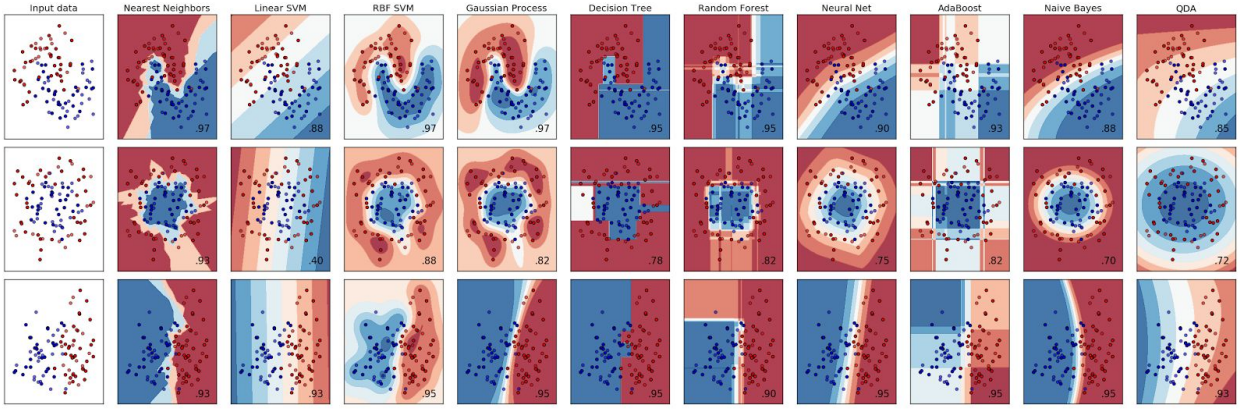
With Random Forest Classifier, we were able to generate a benchmark of 0.7970. So, first we will start with going to deeper understanding of Random Forest worked and what features contributed it to generate this score in training.

(Basic 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 Neighbors, RF)
    - Parameter Tuning
4. Evaluation. Back to 1 with.

For next submission we are expecting a training benchmark of 88%.

Considered Model: Gradient Boosting Trees, Nearest Neighbors, Random Forest(RF).



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 Tanzanania
- UN Report
- UN 2007 Water Taps Installation
- GBT Video Lecture
- GBT
- Classifier Comparison