Capstone 1: Data Wrangling

Topic: Pump it Up: Data Mining the Water Table.

Task: Predict which water pumps are faulty? (predict 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, 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. A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

Data: Using data from Taarifa and the Tanzanian Ministry of Water, Data was collected using paper reports and feedback via phone calls. The data set has 40 features, like water quality, gps_hight, region code, extraction type and location etc.

The data set has 59400 entries and 41 features including the target column. The target column has three values (functional, functional-needs repair and non-functional). Most of the features are categorical.

What kind of cleaning steps did you perform?

The data set has many similar features like (region_code, district_code, ward, lga, subvillage),(payment,payment_type),(extraction_type,extraction_type_group,extraction_type_class),(source,source_type,source_class),(waterpoint_type,waterpoint_type_group)gives similar information, so the correlation among the similar data are very high and removed those columns to avoid risk of overfitting. All the values in the num_private column is zero so, removed num_private from the data set.

How did you deal with missing values, if any?

The data set contains many zero values (amount_tsh, gps_hieght, population, construction_year and population) and missing values in wpt_name.

The zero values in (amount_tsh, gps_hieght, population, construction_year and population) are filled with mean values grouped by region_code. Due to the huge number of factors and the lack of a clear dominating value droped 'wpt_name' column.

Were there outliers, and how did you handle them?

There are no significant outliers in the data set.