PUMP IT UP: DATA MINING THE WATER TABLE

Hosted by DrivenData.org

MACHINE LEARNING 2 GROUP ASSIGNMENT

team 0-1-5: Rafael Hernandez, Hatem Hassan, Fei Dai, Theo Willems, Mrad Azoury, Emily Fuller and Muhammad Furqan

Objective

Water is an essential element of all life on the planet, and access to reliable water services is integral to human health. This service can be taken for granted, yet it is important to account for the infrastructure that needs to be monitored and maintained by governments in order to facilitate access to such a basic utility. The following report will analyze water pumps across Tanzania, with the purpose of developing a classification model that can determine the functional status of each pump. It is off the back of this analysis that a brief operational plan for the maintenance and overall management of water resources will also be provided.

Data Understanding

The provided dataset contains a set of 59,400 values on the training set and 14,850 values on the test set, for a total of 74,250 water pumps spread through the country of Tanzania. For each record, there are a total of 40 variables is given, which describe a series of operational and geographical variables that characterize the functionality of the water pump.

The 3 target values pertaining to the functional status of the pump are as follows:

1. Functional (54.3%)
2. Functional needs repair (7.26%)
3. Non-functional (38.42%)

To begin analysis of the dataset, each of the well locations was plotted on a map of Tanzania using "Carto". This allowed the team to visualize the set of coordinates that were located outside the map of Tanzania, providing clear evidence of data integrity issues that required a deep dive into all the variables.

After little time analyzing the dataset, it was clear that the biggest challenge associated with this dataset was going to be the interpretation and solving of issues related to the integrity of the data, including duplicate values, misspelling, and missing and invalid values. Deciding how these peculiarities would be treated would impact the future modeling, and were integral to address for adequate data preparation. A summary of insights and criteria that informed the Data preparation required for this dataset.

Numerical Features

· **Amount_tsh** (total static head): represents the amount of water to waterpoint. This variable has 70 % of the records with value of zero.
· **gps_height**: Altitude of the well. Approximately 34.4% of the values missing or at 0.
· **Longitude & latitude**: Coordinates of each well. 3% of the values are located outside the Tanzanian territory.
· **num_private**: With an undisclosed description, this variable is a candidate to be dropped, as the feature is underrepresented, with 98% of its values at 0.
·**Population:** people around the well. 35.9% with value 0, 11.8% are at value 1.
·**Public_meeting**: Undisclosed description. Is a binary variable 85.8% True / 8.5% False / 5.7% missing values.
·**Construction_year**: 34.8% with value 0. It is important to note that some of the values for variable date_recorded are prior to the publish date of the dataset, indicating an issue with these values.

Categorical Values

· **Funder:** Who funded the well. A total of 1898 unique values where the top 10 values represent approximately 38% of the population. 6.11% missing values.

·**Installer**: Organization that installed the well. With 2146 unique values, the top 5 values represent close to 38% of the population. 6.15% missing values.

·**Wpt_name**: Name of the waterpoint if there is one, for a total of 37,400 unique values with 5.9% missing values "none".

·**Geographical variables: Basin** (9 unique values)**, District Code** (20 unique values), **Region** (21 unique values), **Region_code** (27 unique values), **lga** (125 unique values), **ward** (2092 unique values) and **sub-village** (19288 unique values). These geographical variables were ordered in decreasing order of granularity. Since these are subdivisions of each other, "cramers_corrected_stat" function was used to evaluate the similarity between then for overlapping or duplicate fields

·**extraction_type** (18 unique values), **extraction_type_group** (13 unique values), **extraction_type_class** (7 unique values): all fields proved to have a high similarity of values. 99% of records confirmed as "matching" using "cramers_corrected_stat "similarity test".

·**management** (12 unique values) and **management_group** (5 unique values): Both variables have the same description "How the waterpoint is managed" and a 99% similarity of records according to "cramers_corrected_stat".

·**payment**(7 unique values) and **payment_type**(6 unique values): Described as "What the water costs" at the well, these variables have both the exact same amount of unique values as well as a 100% matching pattern.

·**water_quality** (8 unique values) **& quality_group** (6 unique values): Described as " The quantity of water" at the well, these variables have almost the same amount of unique values as well as a 100% matching pattern.

·**quantity** (5 unique values) **& quantity_group** (5 unique values): Described as "The quantity of water" at the well, these variables have both the exact same amount of unique values as well as a 99% matching pattern.

·**Source** (12 unique values), **source_type** (12 unique values) & **source_class** (12 unique values) : Described as "The kind of waterpoint" at the well, these variables have both the exact same amount of unique values as well as a 99% matching pattern.

·**waterpoint_type**(7 unique values) **& waterpoint_type_group** (6 unique values): Described as "The kind of waterpoint" at the well, these variables have both the exact same amount of unique values as well as a 99% matching pattern.

Data Visualization

Data visualization can be a helpful tool for exploring the data, and visualizing findings. Given the dataset provided contained geolocation data, "Carto" was useful for mapping the values for further visual analysis and spread of the data geographically. The Carto maps facilitated the identification of outliers, and also allowed the user to visibly identify groupings of pumps by *status_group* throughout the map. **This finding led the team to investigate and add a cluster analysis as a feature engineering variable.**

In addition to the maps, a distribution bar chart was plotted for every categorical variable against the target variable. These allow the user to identify clear indicators against the target variable and to identify issues related to the data. As an example of a clear indicator against the target variable, **pumps with a *quantity* variable value "dry" are a clear indicator of a non-functional status.**

Water_quality, quality_group "unknown" values, and "other" waterpoint_type values were also identified as highly correlated with non-functional water pumps.

A plot of the distribution *amount_tsh* against the target variable also yielded insights pertaining to data integrity. This variable describes the amount of water at the head of the pump, therefore, it is unlikely that many *functional* pumps would be have an *amount_tsh* value of zero, which was indeed the case with the given data. A plot of this against the target suggested unreliable data and an opportunity for additional data treatment, to be described in further sections. In the case of variables *construction_year* and *extraction_type* a relationship was found between the type of extraction (probably associated with technology availability) and the construction year. This relationship proved to be helpful on the imputation of *construction_year* missing values during data preparation section.

Data Preparation and Feature Engineering

Imputing:
In order to impute the missing *gps_height*, each missing value was either replaced by the mean of its corresponding *subvillage*, or another action was taken. If this information was not available it was otherwise replaced by the mean of its *ward*, *lga*, *district_code, region* and *basin*. This was performed in order to ensure that the highest level of granularity possible was used when imputing these values. As each value represented as 0 are considered to be missing, these were first transformed to NaN guarantee that the mean of each group was calculated by neglecting these values.

The same imputation method was used for the variables *longitude*, *latitude*, *population*. Although, latitude and longitude missing value were considered to be all value with -00000002 and not 0.

Regarding the variable *amount_tsh*, the same process for imputing could not be taken for all missing value pertaining to the following 4 regions; Dodoma, Kagera, Mbeya, Tabora. For this reason, the total mean was instead used as no more granular information was available to impute theses missing values.

Conversion:
In order to give more substantial numerical meaning to the variable *contruction_year*, the variable *age* was added. This variable calculates for each row the number of days that elapsed from the construction to today. The missing values in *age* were then imputed using the same method for longitude, latitude, population and gps_height.

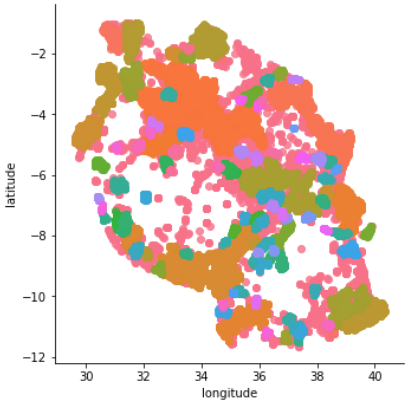
Feature engineering:
A new variable named *density* was added. This variable is the result of dividing the population of its respective row by the total density of its region. The total density of the region was collected from external data available at http://statistics.go.tz. This variable was created for the purpose of complementing the relationship between the geographical space and the amount of population.

Additionally, a new variable *distance* was created to calculate the distance between each pump's geographical position and the geographical position of the capital. In order to do this, the haversine distance formula was used to calculate these distances. As the earth is a sphere, simply using Euclidian or Manahan distance would not grasp the true distance between points. The reasoning

behind this feature engineering method comes from the idea that, the further away a pump is located from the capital, the less it is exposed to infrastructure and maintenance. This may affect the functional status of the pump, as it is less likely to receive routine or needed maintenance.

Following the logic of creating feature from external data, *PTR* was subsequently added. *PTR* denotes the pupil/teach ratio for the region of each row. This variable was likewise collected from one of the many databases available at http://statistics.go.tz. Adding PTR would add a sense of wealth and education within a region which could arguably impact the treatment of pumps and affect the status of the aforementioned.

Finally, a clustering technique using DBSCAN was performed in order to potentially identify geographical groups of pumps that tend to behave similarly. This analysis resulted in classifying all rows in 101 different cluster across Tanzania.



Shortlisting

Shortlisting allows for the elimination of noise for those variables with a widespread distribution across several unique values. As an example, variable *installer* was shortlisted to the top 5 values representing the top 35% of the population, while the rest were labeled as *other*. Note that this process was repeated for variables *funder, lga, extraction_type, scheme_management*, and *region_code.* In addition to eliminating noise, shortlisting allows the user to standardize the train and test set with the same amount of values per variable before being encoded.

One_hot_encoding vs Label_encoder

Different models require different data preparation depending on the algorithm. For the purpose of the modeling section, both One_hot_encoding and Label_encoder were performed to different iterations of the dataset, to meet the requirements of the model and also for the efficiency of the training process. In the case of Logistic Regression, One_hot_encoding was performed, while Label_encoder was performed for Tree-Based models such as RandomForest Classifier and Extreme Gradient Booster Classifier.

Efforts that failed to add value

Throughout the data preparation and feature engineering process the team made an effort to test different variables and features that would enhance the performance of the model but failed to add any values. Some of the features include.

- **SMOTE**: aimed to fix the under representation of “Functional needs repair” from the. target variable.

- **PCA/LDA:** Dimensionality reduction of the variables.
- **Elevation:** Impute the gps_height missing values with an external elevation database using the longitude and latitude coordinates.

Modeling

Different models were tried in order to see which one will coincide best with the problem at hand. Since the data is mostly categorical it made sense to start with **tree based model**s such as **Random Forest** because of the way it splits data into the tree multiple categories helping the model perform better.

K-Nearest Neighbors was trained on the latitude and longitude of the wells to examine any relationship which may exist between the location of the wells and the target.
Logistic Regression was evaluated on the whole dataset, and the results were not fruitful due to the large amount of categorical data and the nonlinear relationship between the different categories and the target.

Baseline (Internal Scores)

- Random Forest: 0.8121212121212121
- XGBoost: 0.7455387205387205
- Multiclass Logistics Regression: 0.7207912457912458
- KNNeighbors: 0.680050505050505

Final Model:
Following data analysis, preparation, feature engineering and modeling, it was clear that Tree-Based models would be ideal for the given dataset. Additionally, performing gridsearch on Random Forest and XGBoost was necessary in order to extract the most value from these models to yield their most optimal parameters.

Grid search was performed with a cross-validation on both models to ensure no overfitting and that the result obtained was valid. The respective best parameters are used from this point on.

Once parameters were optimally tuned, multiple combinations of the data preparation and feature engineering ideas were utilized to yield the optimal combination. This process was iterative, allowing the model to learn the most about the target from feature combinatorics.

Eight different combinations of data preparation were tried with cross-validation using the tuned RandomForest each time getting predictions on the test set. All the mean scores were very close, ranging from 0.81 to 0.82. The two highest scores were scored externally on the Pump It Up competition and gave 0.8243 and 0.8240.

Top 10 - Feature Importance

Rank	Variable name	Weight
1	quantity	0.082
2	quantity_group	0.077
3	longitude	0.064
4	latitude	0.058
5	distance	0.058
6	date_recorded	0.039
7	wpt_name	0.038
8	subvillage	0.037
9	construction_year	0.036
10	waterpoint_type_group	0.035

Two stacking approaches were followed:
-Using all the previous datasets of combinations and predicting on the train for each fold using cross-validation, then on the test set. A new XGBoost model was trained on the 8 train set predictions and test on the test predictions. This method was used to capture different information from our features and to understand how the model interprets them, but
returned a score of 0.8175 on the website.

- Predicting with the dataset that has the combination that resulted in the best score using Random Forest and XGBoost. These predictions were made on the train set then stacked. Another XGBoost was trained for predicting the test set. This approach was used to capture the different results from the two models, and to increase the score. This attempt led to a lower score.

Additionally, different combinations of feature engineering were tried: isolating each feature engineering combination manually and training the model against it both for an internal accuracy score as well as for the Web Page submission.

Applications and Conclusions:
The prediction of water pump status has applications to both the governments and citizens and Tanzania as well as other demographics. This can be useful for the Tanzanian government to prioritize initiatives for increasing access to clean water. Understanding the geographical locations and other variables which correlate with dysfunctional pumps allow for optimal and predictive maintenance targeted towards these areas. Understanding these factors can also help the government to improve placements of new pumps wherein the environment is most favourable (accounting for suitable water quality, for example). This can help the government to mitigate factors which can yield poor performance over time by adapting materials used for water pumps where quality is not-favourable,

such as adding corrosion-resistant materials such as copper, aluminium, bronze or brass to old or newly-placed pumps.

The model can be evolved to include additional data from the Tanzanian government, and should be maintained over time to account for changes in the environment. Maintaining a model which can predict pump status can have long-term benefits for public health by increasing access to clean water and decreasing unnecessary health costs due to dehydration. This model can help other areas to understand important factors for pump functionality, and should be considered for the objective of increasing access to clean water internationally.