CAPSTONE PROJECT: PUMP IT UP

Data Analysis with R

Tadaaki Sun
Foundations of Data Science

Capstone Project Report

Introduction

Tanzania is one of the poor nations around the world who is currently suffering from issues that are related to the water situation. Approximately a third of their country is dry and rarely receive rain, where they have minimal access to water that is clean and drinkable. There are several water points that provide water to the community of Tanzania, however, they are usually far away from villages and hard to retrieve. To solve the many arising issues related to water, the Tanzanian Ministry of Water has decided to hire TDS (Tadaaki's Data Science) Consulting, to have assistance in predicting which water pumps in their country are either, non-functional, functional, or functional but needs repair.

Through several meetings with the Tanzanian Ministry of Water, TDS has observed that the Ministry of Water are interested in understanding the core reason behind the fault of water points in Tanzania. To specifically find the reason for default, any waterpoints that are functional but needs repair will be counted as non-functional for simplicity.

As the Tanzanian Ministry of Water and Taarifa thrive to provide water throughout the Tanzanian community, we would like to propose a recommendation based on our analysis using data analysis tools. The recommendation will then be used to tackle the root issue of the water issue and plan to prevent in the future by providing clean and accessible water points for the Tanzanian community.

Objective

The summarize the objective of this resport, we have identified the general and specific objectives being covered:

- **General Objective**

Use statistical methods to predict the root cause of non-functional water points, which will further help in improving the maintenance operations in water points in Tanzania to ensure that clean, and portable water is available to the Tanzanian communities.

- **Specific Objectives:**

  - Apply exploratory data analysis and visualization techniques to gain insight into the relationships between the different variables in the dataset.
  - Determine which variables are the most associated with the different status of the water points.
  - Create a predictive model that can be used to predict the status of water points (functional, non-functional, and functional but need repair).
  - Create a list of actionable recommendations from the analysis of the data and the results of the model.
  - Provide deliverables (report, presentation slides, and code for the model) to assist

The code, report, and presentation slide which include the detailed analysis and predictive model used to identify the recommended solutions for the Tanzanian Ministry of Water. We hope the documents can be a guide to predict more issues and bring a great community in Tanzania.

Taarifa Dataset

The Tanzanian Ministry of Water has partnered with Taarifa, who currently is working on an Innovation Project in Tanzania, who provided the data such as the status of the pump, when they were installed, how they are managed, etc. They have gathered them from hand held devices, and papers to be gathered in a excel sheet.

The data consists about 40 features and 59,400 observations as a training set, and a further 18,500 observations for the test set, that may assist us in creating a predictive model for finding the non-functional waterpoints in Tanzania. The model will be used in a test dataset provided to measure the effectiveness (the full details of features are in the Appendix section).

In addition to the features above, each water point id will have information on which status group (functional, functional but needs repair, or non-functional) it belongs to. This data will be the dependent variable which will be tested with the other features to predict the reason the pumps are not working.

For our own purposes, we will treat any waterpoint status of "functional but needs repair" as "non-functional" status for simplicity in identifying the reason for default. Furthermore, we will use only the training set to create a predictive model to check the efficacy of the model for future predictions.

The rest of the variables are the independent variables used to investigate what the root cause of defaults of the waterpoints. Most of the data are categorical variables, where there are nominal and ordinal, and continuous variables. Some of the nominal categorical variables include the installer/funder of the waterpoint, information of region/village/etc., and source types. Ordinal categorical variables include the availability of public meetings, permits, etc. Lastly, there are continuous variables such as the amount of water heads, population, and construction years. With a great amount of data for each feature, a thorough preliminary data analysis was done to better understand the data.

Base Line Performance

To identify a feature which is the root of the cause for the faulty water points, the final predictive model will be classified as either functional or non-functional. The amount of water points that were functional but needs repair was only a fractional amount in the whole training dataset.

Once the status is changed, we have a result were there were 32229 functional and 27171 non-functional waterpoints. Therefore, the current data will have approximately 54% (32229/59400) accuracy if we did not have a model. This is not a good predictor as there is only a half chance of guessing the outcome. Therefore, we will create a new model that will have a higher accuracy compared to 54% to better identify the non-functional waterpoint.

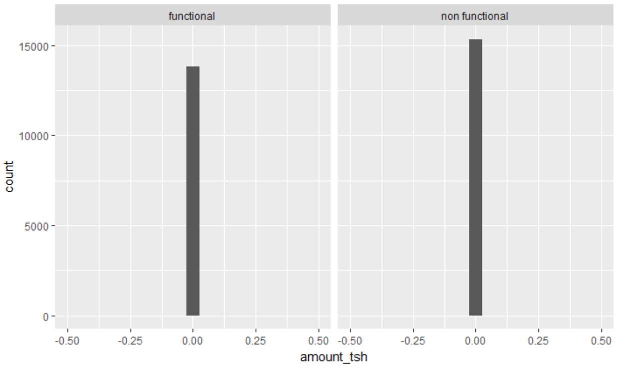
We will then split the data into a training and testing set by having 70% of randomly selected data in the training and 30% in the test set. The new datasets also calculated a result of 54% (22582/41581) accuracy for the baseline performance.

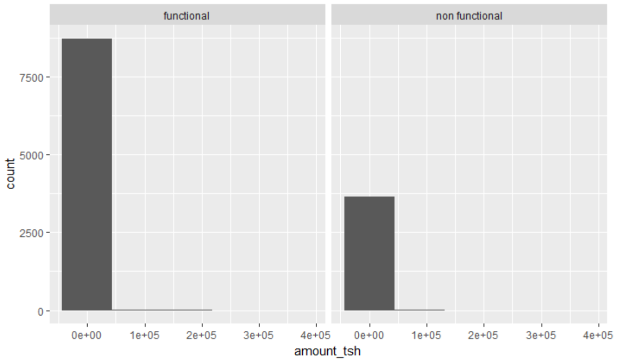
Preliminary Data Analysis

Once the baseline performance is set, the next step is to examine the independent variables. All features were carefully examined by utilizing various visualization techniques such as bar plots and histograms to make sense of variables such as quantity, quality group, water point type. Furthermore, any missing plots that may cause the graphs to skew can be identified at an early stage. We will adjust the data to have a better distribution to create variability in each feature. The utilization of various R packages can assist in creating better visual diagrams and apply some statistical classification methods to have a better prediction with the data. For example, the ggplot2 package can will be utilized to create bar charts and histograms that will give us insights into features that are currently non-functional.

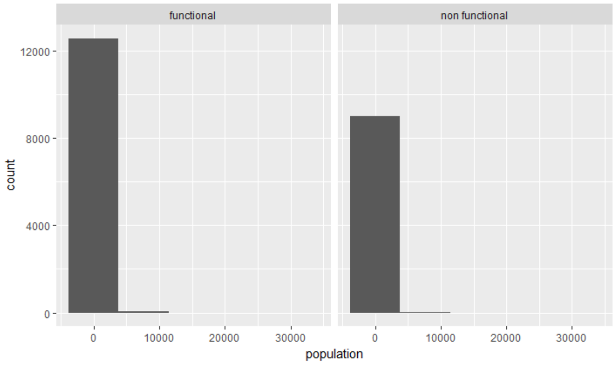
With the first preliminary exploration, there were a lot of missing data as the information were coming from various sources such as handheld devices to paper. There are various levels of features which provided similar results and there is a need to categorize them as duplicates. Most of the operation data had thousands of high levels and are hard to predict the main cause. The training data had informed us that approximately half were functional and non-functional and a fraction being functional but needed repair. Finally, there were few old water point information and a lot of recently implemented pumps. These findings are all limitations when creating the predictive model and would need to be carefully assessed.

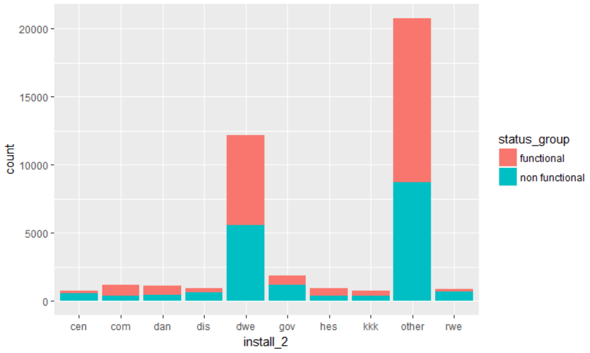
By quickly observing the data in the preliminary data analysis, there were many variables where the information was missing or logically questionable about the results. The dataset provided by Taarifa does a good job in tracking most of the variables with minimal NA’s, however, there were many 0’s and unknowns that still hinder us from providing insights. For example, there is one feature, “amount_tsh”, which is the amount of water around the waterpoint. There are nearly 30,000 waterpoints, yet half were functional or non-functional (left hand diagram below where water amount equals 0). In a logical perspective, a waterpoint without water would mean that it can not provide any water to the community around, and therefore should be characterized as non-functional.



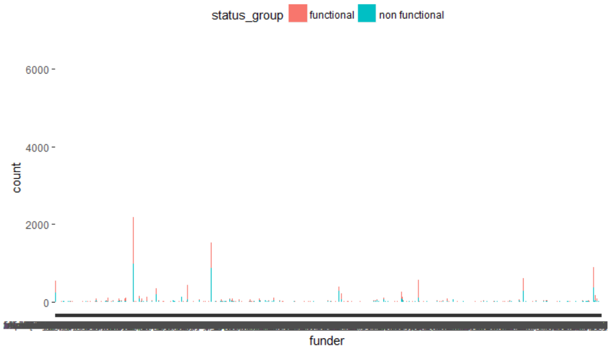


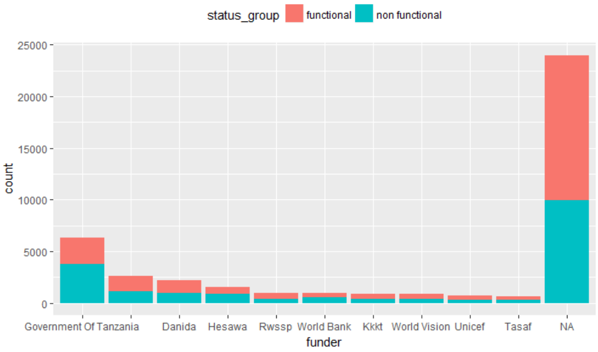
A further analysis was done to check the status for waterpoints that at least have more than 0 water (right hand diagram on the previous page). By creating a histogram for waterpoints that had water, two-thirds of them were functional and one-third was non-functional. There is a good case where the amount of water could be an amazing identifier for the faults, however, there were too many waterpoints that did not have water, but functional. Therefore, we decided to remove this and any other feature (for example population, indicated in the graph below) that had a lot of poorly recorded data to find a feature with correct data.





Furthermore, features where majority of the results were "other" or "NA" were removed, as we are not sure the exact reason why the water pump is at fault. The installer feature had one majority data, despite the "other's", however, the rest of the data were too small. The funder feature had over 2000 levels, where the majority were NA's once grouped. It is unfortunate that these variables are ignored for the purpose of the model.





Risks and Limitations

Before we move into the exploratory data analysis, we will need to identify the changes we will make to the data, and what types of risks and limitations there are. Based on the preliminary data analysis, there are few risks and limitations with the data:

1. The missing values will either be omitted by logic and assumptions – there may be 0's and NA (not available data)'s resulting from mistakes and no proper methods of collecting certain data
2. The duplicate features will be removed – features that are strongly correlated does not add value in the model

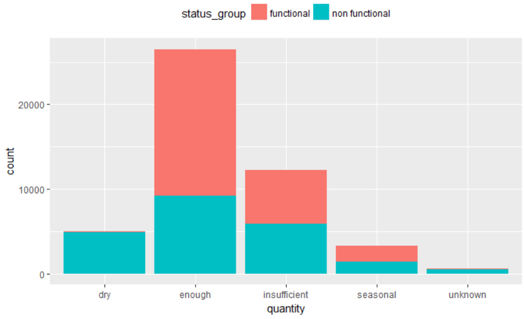
3. Features with high levels of data will be narrowed – for example, there are a lot of installers and funders, where we simply can't visualize well
4. Assess the baseline performance to use as a benchmark when creating the predictive model – as there were no models created in the past, we are assuming this will be the best accuracy today
5. Features with a good amount of data will only be considered in creating a predictive model

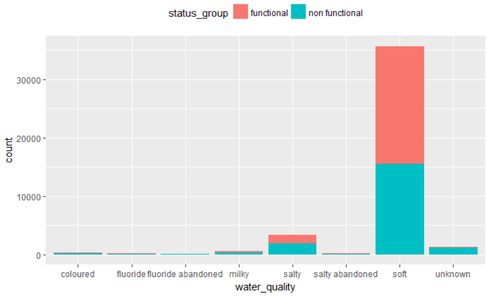
Once the data is cleaned and ready for further analysis, various classification models (logistic regression, random forest, generalized boosting regression model, XGBoost, SVM) will be considered for use to find the nest model for this case. The best model will have a significant increase compared to the baseline performance A feature with higher coefficients would have a strong relationship with the dependant variable, and therefore would be the cause for the non-functional waterpoints.

Exploratory Data Analysis

After visualizing and going through all features that have a significant amount of data for the feature, we decided to compare them to reduce the amount of noise when creating the model. The features that had the most significant data were region, public meeting, permit, basin, payment type, management group, extraction type group, waterpoint type, GPS height, quantity, water quality, scheme management, and source. The following features has minimal amount of missing data and so we decided they were significant enough to contribute to the model.

Some important feature we would like to highlight are ones where we did not have to modify and those we had to reduce or group the levels of information. Quantity on the bottom left hand side is an example of a feature that has enough variability where no change was made. On the other hand, water quality on the bottom right hand side needed an adjustment for the model. We have created a new feature where we group the information based on whether it was soft or not. We can clearly identify if this was significant rather than having one outlier skewing the result, if the model was run with no adjustments.





With the features decided, we carefully looked at the model that makes the most sense in predicting the accuracy of non-functional waterpoints. In the beginning, the methods considered were logistic regression, random forest, SVM, or XGBoost. After carefully examining how different predictors calculates the accuracy, logistic regression was the simplest algorithm that can predict accurately with lowering the noise of the features.

The first step was to remove features such as latitude and longitude, which are the coordination points in a map. For a logistic regression model, it would see the numbers as significant, when there is no meaning in the numbers specifically. After removing them, any outliers/ data with high levels were grouped so that the noise of the model is reduced. Then the features were used in the logistic regression function. The features with less significance (higher p-value), were removed from the model and resulted with a prediction model to be tested with the current training and test data sets.

Final Results

The model that consists of region, public meeting, basin, permit, group of payment type, group of extraction type, group of GPS height, waterpoint type, quantity, group of scheme management and source, resulted with an accuracy of 72% with the test set and 69% for the test set. This model predicted the status groups of the waterpoint with more than 18% accuracy compared to the baseline performance. Not only it provided us with a better prediction, but the coefficients of the features, informs us that quantity, region, and waterpoint type are a good predictor to why the waterpoints are at fault. With this predictive model, we would like to inform Taarifa and the Tanzanian Ministry of Water the next steps.

Recommendation

For our recommendation, we would like to propose a solution that will LIT up the community of Tanzania. We would like you to learn(L) from your mistakes, improve(I) on the current data, and track(T) new data to improve the model and create a better data hygiene to figure out insights.

Taarifa and the Tanzanian Ministry of Water will learn from the current model, where you will locate any waterpoint that are dry, regions that produced high coefficients via the predictive model, and waterpoint type that was classified as “other”. These were the most significant reasons for the default of the waterpoints. The following areas will be a starting point find out the faults and further recorded.

Improving on the current data is also a critical task as there are useful features to include in the model. I have introduced the water mount near the waterpoint which is a feature that could have been used, if the amount of water was recorded. Population is another feature to be considered in the next model as well. There is no meaning in having a waterpoint where there are 0 population. This information needs to be properly collected to be used in the model. Lastly, the funders and installers will create a lot of noise in the model with the amount of high levels. The government can use the model the predict which of them creates the best waterpoint type that will last long to supply the citizens.

To expand on learning and improving the features and models, there is also the need of tracking new information that will make positive additions for the future. While accessing the waterpoints that need to be fixed, I would recommend implementing new data features that will make a difference in future models. The recommendation is to record the failure reasons of the waterpoints, track the remove data to find the average lifetime use of a waterpoint, and track the shortest distance for closest and alternative waterpoints.

In conclusion, there may be a lot of improvements to finding a solution for the non-functional waterpoints. However, this report includes the first analysis of the key problem that is causing the issue, and further recommend the next steps to quickly learn more about the waterpoints. With a better data hygiene and keeping in record of the data with few mediums, the data would be clean and more efficient. Please take the time to LIT up your country by learning from these insights, improving the features, and tracking new information that will help you quickly identify the next faulty waterpoint and have a brighter future.

Appendix

Full Data Features

- amoun_tsh - Total static head (amount water available to water point)
- 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 water point 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 water point is permitted
- construction_year - Year the water point was constructed
- extraction_type - The kind of extraction the water point uses
- extraction_type_group - The kind of extraction the water point uses
- extraction_type_class - The kind of extraction the water point uses
- management - How the water point is managed
- management_group - How the water point 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 water point
- waterpoint_type_group - The kind of water point