Tadaaki Sun

Foundations of Data Science

CAPSTONE PROJECT MILESTONE REPORT: PUMP IT UP

Data Analysis with R

Capstone Project Milestone 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 Ministry of Water, TDS has observed that they are interested in understanding the core reason behind the fault of water points in Tanzania. As they 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.

The details of the objective are specified below:

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 to assist in

TDS Consulting will deliver 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.

The data consists about 40 features (listed below) that may assist us in understanding the bigger picture of the problem to create a predictive model. The model will then be used in a test dataset provided to measure the effectiveness.

Provided 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

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.

Data Wrangling/Preliminary Data Analysis

The intended approach to the problem is by first organizing the data. The trained data will need to be merged (labels and values) to figure out which water point is functional, non-functional or functional needs repair. Various visualization techniques will be used such as such as bar plots, histograms, and scatter plots to make sense of variables such as quantity, quality group, water point type, etc., to have a better picture of the cause of non-functional water points. Furthermore, any missing plots that may cause the graphs to skew can be identified at an early stage. 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, googleVis package can identify ways to visually represent the data and Random Forest can help create the best predictive model.

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 variables 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.

Based on the findings:

1. The missing values will either be omitted or filled by assumptions
2. The duplicate values will be removed
3. Features with high levels of data will be narrowed
4. Assess the baseline performance to use as a benchmark when creating the predictive model
5. Interpret the dates in a meaningful feature

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 used and compared to find the nest model for this case.