TANZANIAN WATER WELLS PREDICTION

DSF-FT2 Phase 3 (Project)

Members

Faith Makokha
Susan Mungai
Willy Angole
Abduba Galgalo
Femi Kamau
John Mungai

ISSUED BY

Moringa School

1. Business Understanding

1.1 Overview

Water makes up about 70% of the earth's surface and is among the few basic needs common to all living creatures. Unfortunately, 96.5 percent of its coverage makes up oceans, which makes up for some scarcity on land.

Tanzania is one of those countries that have to grapple with the scarcity of this life saving need. As a developing nation with a population of over 57,000,000, it has a hard time meeting the demand for safe drinking water due to limited resources for water extraction. There are already some water pumps available in the country but unfortunately, not all are functional and others need repair.

The World Health Organisation- (WHO), understands that three out of ten people still lack access to clean water, highlighting the critical need for investment to improve water hygiene and accessibility. Water poverty has been considered to drive serious illnesses, high infant mortality rates, poor education, slumped economies and unproductive agricultural conditions in a majority of the regions.

1.2 Problem Statement

Tanzania is currently failing to meet the potable water requirements for its population. Despite already building water wells to offset this problem, many are unfortunately non-functional and others are in need of repair.
As a result, our stakeholder, WHO, has narrowed down their interest to drilling more water wells and maintaining/ revamping existing ones within the country, in an effort to ensure availability of quality and quantity drinking water countrywide .

Our role as the data scientist in this project will be to identify patterns in non-functional wells, with the aim of influencing how new ones are built. Furthermore, using these patterns, we will enable our stakeholder to accurately predict existing water points in need of intervention ensuring the people of Tanzania have access to clean potable water.

1.3 Project Justification

Climate change, increasing water scarcity and population growth pose challenges for water supply systems. The World Health Organisation (WHO) has a mission to ensure quality drinking water to water-stressed countries, including Tanzania. They want to build wells in Tanzania, ensure the sustainability of the new ones and maintenance of old ones in the country.

1.4 Specific Objectives

- To identify the trends/patterns between both non-functional and functional wells
- To identify non-functioning wells using a simple analysis, and predict the functionality of a well based on available variables.

1.5 Research Questions

- How can we apply Machine Learning and necessary classification methods to predict functionality of wells in Tanzania?
- Are there any common features that are consistent with functional and non-functional wells?
- Can we identify ways to help build better wells?

1.6 Success Criteria

1.6.1 Business Success Criteria (Review)

- To ensure that newly constructed wells are of good quality water for the communities.
- To correctly identify functionality of a well and determine its viability.
- To successfully predict the type of wells that dry up and determine the type of wells to build based on the area?
- To determine working wells in different regions and which ones need repair or improved management.
- To ensure built wells are longer lasting/ more viable.

1.6.2 Project Success Criteria

Generating a model that will be able to correctly predict the quality status of the wells in Tanzania with an accuracy of 75%.

1.7 Project Plan

Our project will consist of:

- Cross-Industry Standard Process for Data Mining (CRISP-DM) will be used for conducting this research.
- A GitHub repository
- Presentation slides for the project
- A Trello Board for collaboration and tracking

1.8 Scope of the Study

This study explores machine learning models that use datasets obtained from an open-source platform in order to analyse the factors that affect the conditions of water wells.

2. Data Understanding

2.1 Overview

In this project we shall use a dataset containing information about existing water wells in Tanzania sourced from an ongoing DrivenData competition.

2.2 Data Description

The dataset contains 59,400 records and spans 40 columns. Of these columns, we identified 31 to be categorical, and 9 as numerical. We were able to further group the columns into the general features being captured. These were:

2.2.1 Location (Location-based data)

Numerical

- longitude - GPS coordinate
- latitude - GPS coordinate
- gps_height - Altitude of the well

Categorical

- region - Geographic region
- region_code - Geographic location (coded)
- district_code - Geographic location (coded)
- lga - Local Government Authority
- ward - Geographic location
- subvillage - Geographic location
- basin - Geographic water basin

2.2.2 Date (Time-based data)

- construction_year - Year the waterpoint was constructed
- date_recorded - The date the row was entered

2.2.3 Monetary (Monetary-related data)

- payment - What the water costs
- payment_type - What the water costs

2.2.4 Technical (Technical data about the water points)

Numerical

- amount_tsh - Total static head (amount water available to waterpoint)

Categorical

- extraction_type - The kind of extraction the waterpoint uses
- extraction_type_class - The kind of extraction the waterpoint uses
- extraction_type_group - The kind of extraction the waterpoint uses
- 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
- status_group - The condition of the wells (target variable)

2.2.5 Non-Technical (Non-technical data about the water points)

Numerical

- population - Population around the well (those living near it)

Categorical

- installer - The organisation that installed the well.
- funder - Who funded the well.
- wpt_name - Name of the waterpoint if there is one
- public_meeting - True/False
- scheme_management - Who operates the waterpoint

- scheme_name - Who operates the waterpoint
- permit - If the waterpoint is permitted
- management- How the waterpoint is managed
- management_group - How the waterpoint is managed

2.2.6 Miscellaneous

Numerical

- num_private

Categorical

- recorded_by - Group entering this row of data

2.3 Data Quality Verification

After previewing the dataset and observing the quality across the columns, we deduced that we need some detailed cleaning to ensure completeness and consistency. Our dataset is valid and will help in solving our business problem.

3. Data Preparation

Overview

In line with our objectives, where we want to predict wells that are functional, non-functional and that need repair. This dataset's analysis and classification will help our stakeholders to improve maintenance of the present water wells or give useful information for future wells.

3.1 Data Selection.

There are a lot of repetitive columns in the dataset, for example. `` `payment` `` and `` `payment_type` ``,`` `source` `` and `` `source_type` `` which have the same information. We scrapped the columns and for the sake of preparing our data for modelling purposes. We have included columns that contain information about the location and the basin that is closest to the water wells geographically.

3.2 Data Cleaning

The data was checked for missing values and a number were found in the columns; `` `funder`, `installer` ``,`` `subvillage` ``,`` `scheme_management` `` and `` `permit` `` the percentage of missing values in the particular columns were a small amount, thus were dropped to ensure *completeness* of the dataset.

Checking for *validity* in the data, the dataset was checked for any duplicated values and outliers. The duplicated records were not dropped as it does not mean that they were similar wells but just built under the same project.

We also chose not to drop the outliers, as it did not display erroneous data but will be further looked into in the analysis section.

After cleaning the dataset, we need to bring *uniformity* by formatting and the columns to be readable and easily interpretable.So we defined functions to make these possible.

Moreover, we verified that the values of various columns are *consistent*. The names of installers and funders had numerous instances of the same funder and/or installer, but with misspelt names and/or typographical errors.

3.3 Exploratory Data Analysis

The EDA was done by carrying out univariate and bivariate analysis on the features.

3.3.1 Univariate Analysis

The univariate analysis was split into the examination of the categorical and numerical features. With the general purpose of making data easier to interpret and to understand the general distribution of the data. Observation of the categorical features revealed that the `funder`, `installer`, `local_government_area`, `extraction_type_group`, `management_group`, `payment_type`, `water_quality`, and `quantity` columns were not evenly distributed. On the other hand, the `basin`, `region`, `ward`,`waterpoint_type` columns were fairly distributed columns. As for the numerical columns, we evaluated them by visualising the statistical distribution of the columns.

3.3.2 Bivariate Analysis

The bivariate analysis was done with the aim of understanding the empirical relationship between the features columns including our target column.

3.4 Data Quality Report

After cleaning and preparation, the Validity, Accuracy, Completeness, Consistency, and Uniformity of the Data attest to the data's quality.

Findings

- Areas with higher population have a higher number of functional wells.
- Despite being an Important area, Iringa has a lot of non-functional water points which have soft water.
- Most of the wells funded by the government are non-functional.
- Most of the water points which the central government and district council installed are non-functional.
- The most common extraction type is gravity but second is hand pumps.
- The efficiency of handpumps is less than commercial pumps. It shows that authorities need to focus on pumping.
- It is seen that there are many non-functional water points which belong to gravity (which is a natural force so no need to do anything expensive) as extraction type.
- Some water points which have enough and soft water are non-function but we can check the basin first before choosing an area to build the well.
- The wells which have been constructed in recent years are more functional than older ones.
- Some recently built wells are functional but need repair. This means that if they are not repaired soon, they will soon be non-functional .
- There are a lot of wells which have enough water but are non-functional.

4. Modeling

To begin, we point out that our dataset is massive and that there is class imbalance; specifically, 54.1% of our data is considered to be "functional," while 38.8% is "non-functional," and the remaining 7% pertains to "non-functional wells that need repair." We One-Hot-Encoded the categorical variables and categorised the target.
Since the majority of our information comes from operational wells, maintaining the current imbalance isn't a priority for us.
Our baseline model was LogisticRegression with the StandardScaler as our scaler, just to test how our chosen attributes would perform on a basic level.
We then explored other more sophisticated models like the RandomForest, Decision trees, K-Nearest -Neighbours and XGBoost.

Logistic Regression

- The logistic regression model returned an accuracy of 67%. Though this was not our desired accuracy, it was good for a baseline model and gave us a good starting point for our other models.
- The confusion matrix evaluation showed that the model had a bias towards predicting that a well is functional. Furthermore, we saw that the model had a higher number of true positives and true negatives than false positives and false negatives. This therefore meant that the model was not overfitting.
- The cross validation results returned consistent scores. This ensured that we were not overfitting our data

Decision Tree

- The decision tree model returned an accuracy of 70%. This was a better result that our baseline model (logistic regression). However, despite improving the accuracy, it was still less than our desired accuracy of 75%.
- The cross validation once again returned balanced consistent values across the tests

- The confusion matrix evaluation showed that the model had a bias towards predicting that a well is functional. Furthermore, we saw that the model had a higher number of true positives and true negatives than false positives and false negatives. This therefore meant that the model was not overfitting.

Random Forest

- Our random forest model also returned an accuracy of 70%. This was the same as the decision tree model. This meant that the random forest model was not better than the decision tree model. The cross validation once again returned balanced consistent values across the tests
- The confusion matrix evaluation showed that the model had a bias towards predicting that a well is functional. Furthermore, we once again saw that the model had a higher number of true positives and true negatives than false positives and false negatives. This therefore meant that the model was not overfitting.

K Nearest Neighbours

- Unfortunately our KNN algorithm returned some errors and was not able to run.

5. Evaluation

- We used a pipeline to scale the data and then fit it to the model. We then used cross validation to ensure that our model was not overfitting. We also used a confusion matrix to evaluate the performance of our model.
- Despite the fact that our models did not achieve the desired accuracy of 75%, we were able to achieve an accuracy of 70% which is a good start for our project and is within an acceptable range of +/- 5%

6. Conclusion

The accuracy of both our random forest classifier and decision tree models was 70%. While this is still a good predictive model, we would like to undertake further feature engineering to boost this recall score if we had more time.
We achieved our objectives to be able to predict the functional wells and had an acceptable accuracy score.

Recommendations

- When our stakeholder decides to construct additional wells in Tanzania, they may want to look into Lake Rukwa as a basin area, where there are disproportionately more non-functional wells than functional ones.
- The region of Dodoma has more non-functional wells than functional, this area needs to be looked into.
- Wells permitted to operate tend to be more viable and functional over time than those without.
- The wells that are not paid for tend to be non-functional as they are maybe misused by the public, maybe implementing an affordable payment scheme will help curb this.
- Wells without permits also have a higher chance to be non-functional so this means that our stakeholder needs to make sure that they are permitted to ensure they are suitable for human consumption as well.
- Wells with close proximity to Lake Victoria basin tend to be long-lasting compared to the rest of the basin.