Tanzania Water wells - Project
Business Overview

Introduction

This project aims to create a classifier that predicts the condition of water wells in Tanzania. Tanzania, as a developing country, faces difficulties in providing its population with clean, a water, the population of over 57 million people. Many of the existing water points require repair, have failed completely or are fully working,

Stakeholders: The proposed audience for this solution includes NGOs focused on identifying wells in need of repair and the Tanzanian government, which can use the model to detect patterns in non-functional wells and inform future well construction.

Challenges

The dataset may have missing or inconsistent information, which can affect the accuracy of the model. Cleaning and preparing the data is important to ensure reliable predictions.

Deploying the model in real-world scenarios and maintaining its performance over time can be challenging. We need to integrate the model into existing systems and continuously update it with new data.

Selecting the right information from the dataset and creating useful new features can be difficult. It requires understanding what factors influence well conditions in Tanzania.

It's important for stakeholders to understand why the model makes certain predictions. We want to provide clear explanations so that NGOs and the Tanzanian government can make informed decisions about well repair and new construction.

The model should be able to handle a large volume of data efficiently, making predictions quickly and reliably.

Deploying the model in real-world scenarios and maintaining its performance over time can be challenging. We need to integrate the model into existing systems and continuously update it with new data.

To overcome these challenges, we need to carefully clean and prepare the data, select relevant features, train the model accurately, and explain the e results clearly. Collaboration with experts and stakeholders is important, and we should monitor and improve the model continuously.

Proposed Solution

Certify the dataset is free from missing or inconsistent information by cleaning and preparing the data. Data cleaning is critical to maintain the accuracy and reliability of the model.

Provide clear explanations for the model's predictions to enable stakeholders, such as NGOs and the Tanzanian government, to make informed decisions regarding well repairs and new constructions.

Collaborate with the stakeholders throughout the model development process. Monitor the model's performance continuously and strive for improvements based on feedback and new insights.

Integrate the model into existing systems used by NGOs or the Tanzanian government. Continuously update the model with new data to keep it relevant and improve its performance over time.

Conclusions

the project aims to develop a classifier for predicting the condition of water wells in Tanzania. The goal is to assist NGOs and the Tanzanian government in identifying wells that nonfunctional, functional and not functional, deploying the model and ensuring its long-term performance present challenges. To overcome these challenges, it is crucial to clean and prepare the data effectively, select relevant features, and train the model accurately and give clear descriptions for the model's predictions is important for stakeholders to make informed decisions.

- Working with experts, stakeholders and the locals is key to gaining insights into the factors influencing wells conditions. Continuous monitoring and improvement of the model are necessary to maintain its accuracy over time. By applying these solutions, we can address Tanzania's clean water challenges successfully.

Problem Statement

Tanzania, as a developing country, struggles with providing clean water to its population of over 57,000,000. There are many water points already established in the country, but some are in need of repair while others have failed altogether.

Objectives

1. To build a classifier to predict the condition of water wells in Tanzania.
2. For the NGOs to focused on locating wells needing repair
3. the Government of Tanzania, which aims to identify patterns in non-functional wells to improve future well construction.
4. we will develop a classifier to predict the condition of water wells.

Data Understanding

1. We imported different libraries that helped us to load and explore the dataset

2. We are using the Tanzanian Water Wells dataset, which is to be used by the NGO's and the Government of Tanzania.

3. We used inbuilt Python functions to get various insights of our data for example the shape, describe

4. The dataset provided was sufficient hence didn't source more data

5. Columns in the dataset contained both numerical and categorical variables

Data Cleaning

- Upon examination of this dataset, i realized there were no duplicate

- *In order to build a model that can predict the distribution of labels, the relevance and importance of each feature has to be taken into consideration. Here is some of the reasoning behind dropping of certain features.*

  - wpt_name: This column represents the name of the waterpoint. Since the column name descriptions don't indicate that this information is crucial for predicting the waterpoint functionality, it can be dropped.
  - num_private: The column name description doesn't provide any information about this feature, and it may not have a significant impact on predicting waterpoint functionality. Hence, it can be dropped.
  - recorded_by: This column represents the group entering the row of data. If all the values in this column are the same, it doesn't provide any useful information for prediction. You can check its unique values to verify this. If it has only one unique value, it can be dropped.
  - scheme_name: Since it has a high number of missing values (28166), it may not provide significant predictive power and can be dropped.
  - Sub_village: Although it has a relatively small number of missing values (371), it might not be crucial for predicting the waterpoint functionality. Unless you have domain knowledge suggesting its importance, it can be dropped.

Some features have been kept regardless of whether or not they contain missing values because they may have an impact in the overall functionality of the water well

  - installer: With 3655 missing values, this feature may not strongly correlate with the functionality of waterpoints. However, it will be of importance to establish which installer's wells were found to have most faults. This may not necessarily be used in the model training but it offers us valuable insight as to the track record of the installer.
  - funder: Similar to 'installer', if it has a large number of missing values (3635) and does not seem to be highly informative for predicting functionality, however for analysis it is important to establish which funder had the best working wells.
  - public_meeting and permit: Although they have missing values, these features relate to public involvement and legal permissions, respectively, which could potentially impact the functionality of waterpoints. The missing values was handled by imputing the most common value.

- scheme_management: Although it has a high number of missing values (3877), it might still contain valuable information regarding the management of water schemes, which can affect functionality. The missing values was handled by.

Data Analysis

I plotted a graph waterpoint condition distribution and it showed that water wells constructed from 2007 to 2013 are functioning well as compared to wells in 1960.

A correlation matrix can be used to identify variables that are strongly correlated with each other, and may therefore be important predictors of a target variable. This can help in feature selection for predictive modeling tasks.

the graph shows the amount of money spent on the well depends the how well the water wells are taken care of and how they function the more the amount the more they functionality

Modeling

MODEL

I used logistic regression which showed the precision, recall, and F1-score for each class, and provides the accuracy of the model overall. the model achieved an accuracy of 0.60, with varying performance across the different classes. The model's performance is below average. It performs relatively well in predicting the "functional" class, but poorly in identifying instances that need repair ("functional needs repair" class). Overall, the model has difficulty recognizing instances that require repair, indicating the need for improvement. that's why we use the next model.

The KNN classifier achieved an accuracy of 0.70 when making predictions on the test data. The precision, recall, and F1-score for each class are provided in the classification report. The "support" column represents the number of samples in each class. In summary, the model's overall performance is decent, with better performance for the “functional" and "nonfunctional" classes compared to the "functional needs repair" class. However, there is room for improvement, particularly in correctly identifying instances requiring repair. We used the below model,

The Random Forest Classifier was trained on the training data and evaluated on the test data using the Tanzania water wells dataset. Here is a report on the performance of the Random Forest Classifier: The accuracy of the model on the test data is 0.79, indicating that it correctly predicted the condition of the water wells in 79% of the cases. The precision for the "functional" class is 0.81, which means that when the model predicts a water well as functional, it is correct 81% of the time. For the "functional needs repair" class, the precision is 0.49, indicating that the model has a lower accuracy in identifying wells that need repair. The precision for the "nonfunctional" class is 0.81, suggesting that the model performs well in identifying non-functional wells. The recall for the "functional" class is 0.86, indicating that the model correctly identifies 86% of the actual functional wells. The recall for the "functional needs repair" class is 0.35, suggesting that the model has difficulty in correctly identifying wells that need repair. The

recall for the "nonfunctional" class is 0.78, indicating that the model identifies 78% of the actual non-functional wells. F1-Score: The F1-score combines precision and recall into a single metric. The F1-score for the "functional" class is 0.83, while for the "functional needs repair" class, it is 0.40. The F1-score for the "nonfunctional" class is 0.79.

In conclusion, the Random Forest Classifier shows relatively good performance in predicting the condition of water wells, with higher accuracy, precision, recall, and F1-score for the "functional" and "nonfunctional" classes compared to the "functional needs repair" class. there is room for improvement in identifying wells that need repair.

Recommendations
I recommend

1. There is a need to do further exploration into other features in order to better understand the determinants of house prices.
2. The Government should consider partnering with the Locals and non-governmental organizations in maintaining the water wells.
3. The model in Project Phase 3 can be improved by using a binary classification approach.
4. More feature engineering to be carried out.
5. The government should offer incentives to the NGO to enhance more water wells to be dug for its people for example give tax waives for NGOs that dig water wells and maintain them.

Next Steps

1. The road to the water points should be easily accessible for use and maintence.
2. Revise the models so that it reflects the water pumps, installations so that this will allow for better accuracy in predicting the functional, nonfunctional water wells
3. The analysist can query how to improve in identifying wells that need repair.