PREDICTING THE CONDITION OF TANZANIA WATER WELLS

1. Business Understanding

Problem Statement

Lack of clean and potable water is a major issue in many communities across Tanzania. To address this issue, the Tanzanian Ministry of Water has installed several water wells across the country.

However, not all of these wells are functioning as intended, which results in a lack of access to clean water for communities.

The goal of this project is to build a predictive model that can accurately predict the condition of water wells in Tanzania based on data from Taarifa and the Tanzanian Ministry of Water to predict which pumps are functional, which need some repairs, and which don't work at all.

By doing so, the aim is to improve maintenance operations and ensure that clean and potable water is available to communities across Tanzania.

Research Question

Which classifier model can accurately predict the condition of water wells in Tanzania?

Objectives

Main Objective

To predict the condition of water wells in Tanzania to ensure that clean and potable water is available to communities across Tanzania.

Specific Objectives

- To understand the problem statement and the goal of the project
- To identify the variables that can impact the functionality of water wells
- To determine the target variable (functional, need repairs, or non-functional)

Metric of Success

The model will be considered a success when both accuracy and f1 score are between 0.8 to 1.

2. Data Understanding

Data Source

Data is downloaded from "Pump it Up: Data Mining the Water Table" competition hosted by DrivenData

Data Description

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

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 waterpoint 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 waterpoint is permitted

construction_year - Year the waterpoint was constructed

extraction_type - The kind of extraction the waterpoint uses

extraction_type_group - The kind of extraction the waterpoint uses

extraction_type_class - The kind of extraction the waterpoint uses

management - How the waterpoint is managed

management_group - How the waterpoint 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 waterpoint

waterpoint_type_group - The kind of waterpoint

3. Data Preparation

- Uniformity - Handling Missing Values
- Consistency - Encoding Categorical Variables
- Completeness - Ensure the dataset has no missing values.
- Scaling
- Handling Class Imbalance
- Uniformity - Handling Missing Values

4. Modeling

The Random Forest Classifier model has an accuracy of 0.8126 and an F1 score of 0.8059

In terms of accuracy, the random forest classifier has a higher accuracy score of 0.81 compared to the other 4 models, which had accuracy scores of 0.76 (gradient boosting), 0.77 (KNN), 0.75 (decision tree), and 0.65 (logistic regression).

In terms of the F1 score, the random forest classifier has a higher F1 score of 0.81 compared to the other 4 models, which had F1 scores of 0.74 (gradient boosting), 0.76 (KNN), 0.76 (decision tree), and 0.62 (logistic regression).

Based on these results, the random forest classifier appears to be performing the best among the 5 models.

Hyperparameter tuning and cross-validation is performed because they help to optimize the model performance and prevent overfitting or underfitting.

The randomized search cross-validation helps to perform an efficient search for the optimal hyperparameters and cross-validate the model, which leads to better model performance and robustness.

The accuracy of the model is 0.8229, which means that the model correctly predicted the class label of 82.29% of the test data instances.

The F1 score is 0.8150. A higher F1 score indicates a better performance of the model.

This tuned model has better performance metrics and hence a good model.

5. Challenging the Solution

1. While the model has a high accuracy of 0.82, the precision and recall scores for the 'functional needs repair' class are lower compared to the other two classes. This could indicate that the model is not accurately predicting instances of this class and could potentially lead to misclassification of important information.

2. The imbalance in the number of instances for each class in the training data could lead to the model overfitting to the majority class and not accurately predicting instances of the minority class. To address these challenges, further refinement of the model, such as using methods to balance the class distribution in the training data, or using different algorithms, could be considered.

3. Random Search CV is computationally expensive and therefore not the most suitable model for large datasets.

6. Conclusion and Recommendations

Conclusion

- The best performing machine learning algorithm for this problem was the Random Forest Classifier with an accuracy of 82.29% and a weighted F1 score of 0.815
- The model could be further improved by incorporating more data especially for the functional needs repair class to handle imbalance for the classes

Recommendations

- The Tanzania Ministry of Water should invest in better waterpoint types such communal standpipes and hand pumps
- The Tanzania Ministry of Water should ensure that the extraction type for the wells is mostly through gravity and handpump
- The Tanzania Ministry of Water should ensure that the gps height(altitude of the well) for most water points is high enough
- The Ministry of Water should also ensure that the people using the waterpoints pay either monthly, annually or per bucket to ensure that the wells are well maintained