Predicting Water Well Functionality in Tanzania

Exploring ML applications for water wells repair optimizaiton

Insights Consulting LTD
George C

Project Objective

- To develop a machine learning model that classifies the operational status of water wells in Tanzania as functional, non-functional, or in need of repair. This model will then help NGOs and government agencies prioritize well maintenance and repairs, ensuring efficient use of resources and improving water access for communities across the country, in line with SDG6.

Business and Data Understanding

Problem Statement

Operational Challenge: Many communities in Tanzania rely on water wells, but a significant number are non-functional or require repair. It is difficult for NGOs and government agencies to identify which wells need maintenance, leading to wasted resources on unnecessary new well constructions instead of repairing existing ones.

Dataset Overview

The dataset for this project consists of publicly available data from Tanzania's Ministry of Water. It spans from 1960 to 2013, with over 59,000 records of water wells across Tanzania.

- **Operational Details:** Information about the well's management, installation, and funding.
- **Geographic Data:** GPS coordinates, altitude, and region specifics.
- **Water Source and Quality:** Details on the water's source, quality, and extraction method.
- **Maintenance Indicators:** Data on construction year, public meetings, permits, and population served.

Modeling Approach

Step 1: Baseline Logistic Regression Model

Rationale:

- **The baseline logistic regression model was chosen for its simplicity and interpretability. Logistic regression is particularly useful for classification tasks like this one**
- **By starting with a logistic regression model, we established a clear benchmark for evaluating the impact of subsequent improvements.**

Performance:

Mean Cross-Validation Accuracy: 75.89%

- The model demonstrated solid overall accuracy, effectively classifying wells based on the features provided. However, while the accuracy was respectable, we recognized that the model might struggle with classifying minority classes, such as non-functional wells.

Step 2: Hyperparameter Tuning

Method

- **To enhance the baseline model, we applied hyperparameter tuning using GridSearchCV.**
- **This technique allowed us to systematically search for the best combination of hyperparameters for the logistic regression model, including regularization strength (C), solver type, and penalty.**

Performance:

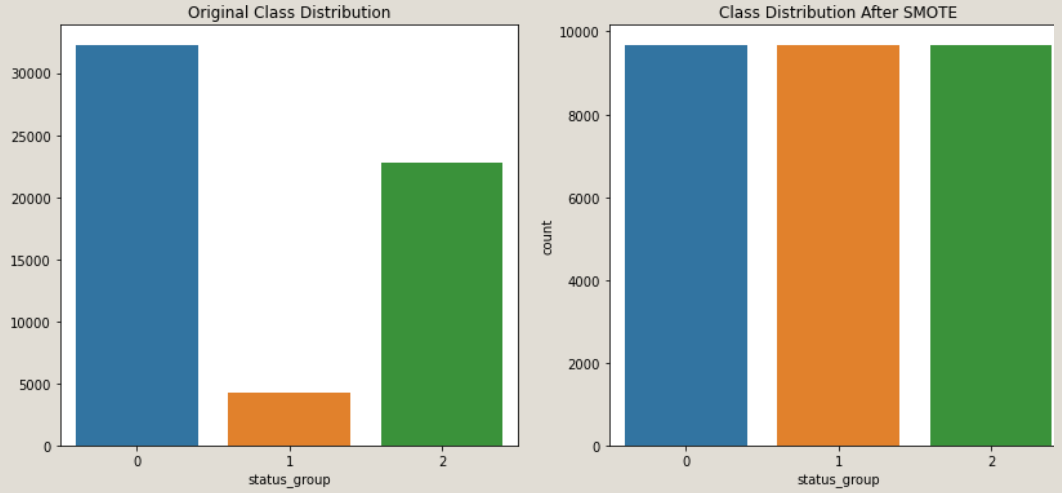
Mean Cross-Validation Accuracy: 75.89%

- The tuned model achieved a slightly higher accuracy, but the improvement was marginal and remained close to the baseline performance, *suggesting that a different approach might be necessary* to achieve a more substantial improvement, especially for minority class predictions.

Step 3: Addressing Class Imbalance with SMOTE

Rationale:

- **One of the critical challenges identified was class imbalance**



Performance: Mean Accuracy: 75.89%

Solution:

- To mitigate this, we applied SMOTE to oversample the minority classes. helping the model to better recognize and predict non-functional wells.

- The accuracy remained on par with the baseline model, but the SMOTE-enhanced model showed **improved recall and precision for minority classes**, making it more reliable for predicting wells that require attention.

Step 4: Decision Tree Model

Rationale:

- **Decision trees were explored as an alternative to logistic regression due to their ability to capture non-linear relationships and provide clear, interpretable decision rules.**

Performance: Mean Accuracy: 74.61%

The decision tree model achieved slightly **lower accuracy** compared to the logistic models. However, its interpretability and ability to handle complex relationships make it a valuable tool for understanding the factors influencing well functionality, even if it did not outperform the logistic regression models in terms of accuracy.

Model Selected: SMOTE Logistic Regression model

Rationale:

- **The SMOTE Logistic Regression model was chosen for its ability to handle class imbalance, which is critical for ensuring that non-functional wells are accurately identified.**

Performance:

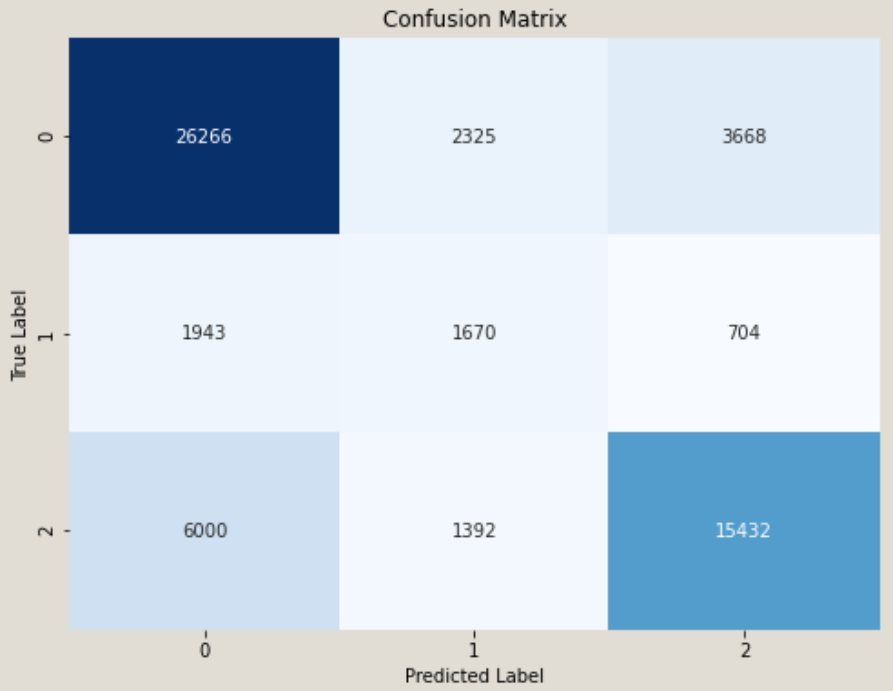
Mean Cross-Validation Accuracy: 75.89%

- The model demonstrated solid overall accuracy, effectively classifying wells based on the features provided. However, while the accuracy was respectable, we recognized that the model might struggle with classifying minority classes, such as non-functional wells.

Final Model Evaluation

Model Evaluation

Confusion Matix



Predicted: Functional (0)	Predicted: Non-functional (1)	Predicted: Needs Repair (2)

Performance:

- The model's main strength is in identifying wells needing repair, which aligns well with the project's goals of maintaining water access.
- However, the significant area for improvement is in differentiating between wells that are functional but need repair and those that are non-functional.
- Correcting these misclassifications will be essential for ensuring that resources are allocated efficiently, targeting wells that truly need immediate intervention versus those that require regular maintenance.

Model Evaluation

Classification Report

Rationale:

Classification Report:

	precision	recall	f1-score	support
0	0.77	0.81	0.79	32259
1	0.31	0.39	0.34	4317
2	0.78	0.68	0.72	22824
accuracy			0.73	59400
macro avg	0.62	0.63	0.62	59400
weighted avg	0.74	0.73	0.73	59400

Performance:

Key Metrics Summary:

- *Functional Wells (Class 0):* Precision: 77%, Recall: 81%—strong overall accuracy.
- *Non-Functional Wells (Class 1):* Precision: 31%, Recall: 39%—model struggles with detection.
- *Wells Needing Repair (Class 2):* Precision: 78%, Recall: 68%—good identification, but some are missed.

Summary: The model performs well with functional wells but needs improvement in detecting non-functional wells.

Business Implications

Business Implications

Resource Efficiency:	High precision (77%) for functional wells minimizes unnecessary repairs, focusing efforts on critical areas.
Missed Repairs:	Low precision (31%) for non-functional wells risks overlooking critical repairs, potentially disrupting water access.
Targeted Maintenance:	The model effectively prioritizes wells needing repair (Precision: 78%), but can be improved for better accuracy.
Strategic Impact:	Enhancing non-functional well detection strengthens efforts toward Sustainable Development Goal 6, ensuring reliable water access.

Business Recommendations

Recommendations:

1. Enhance Resource Allocation: Focus maintenance efforts on wells accurately identified as needing repair to maximize resource efficiency.
2. Improve Detection of Non-Functional Wells: Implement strategies to boost precision for non-functional wells, reducing the risk of missing critical repairs.
3. Align with SDG 6: Strengthen the model's ability to identify at-risk wells to better support efforts in achieving Sustainable Development Goal 6 for clean and accessible water.

Thank You!

We appreciate your time and attention. This project has demonstrated the potential of data-driven approaches to significantly improve well maintenance and resource allocation across Tanzania. By leveraging predictive modeling, we can support the Sustainable Development Goals and ensure that communities have consistent access to clean and reliable water.