Tanzanian Water Wells

Author: Joy Ogutu

Content

1 Introduction

2 Problem Statement

3 Objectives

4 Analysis

5 Modeling

6 Conclusion

7 Recommendations

Problem Statement

Tanzania grapples with water infrastructure challenges, impacting clean water access due to well functionality issues. Traditional wells, limited resources, and climate change exacerbate water scarcity and contamination risks. Addressing these challenges requires data-driven insights, sustainable infrastructure development, and collaborative efforts for equitable resource distribution and community empowerment, crucial for Tanzanian well-being.

Back to Agenda

Objectives

- Explore water quality indicators' relationship with well functionality for targeted improvement initiatives.
- Assess impact of numeric variables (e.g., total static head, population) on well functionality.
- Develop advanced predictive modelformproactive water well maintenanceusing diverse classifiers

Back to Agenda

Data Cleaning and Preprocessing

Duplicates

- There were no duplicate entries in the data.

Null Values

- Null values, except in the "permit" column, were handled by replacing them with "unknown," and date_recorded data type was converted for consistency.

Dropping Columns and Outliers

- Highly correlated and irrelevant features were identified and dropped to prevent overfitting.
- outliers were retained for a comprehensive dataset exploration.

Feature engineering

- Date_recorded data type was converted to datetime for consistency in data representation.
- Feature engineering introduced the "well_age" variable, calculating the age

Back to Agenda

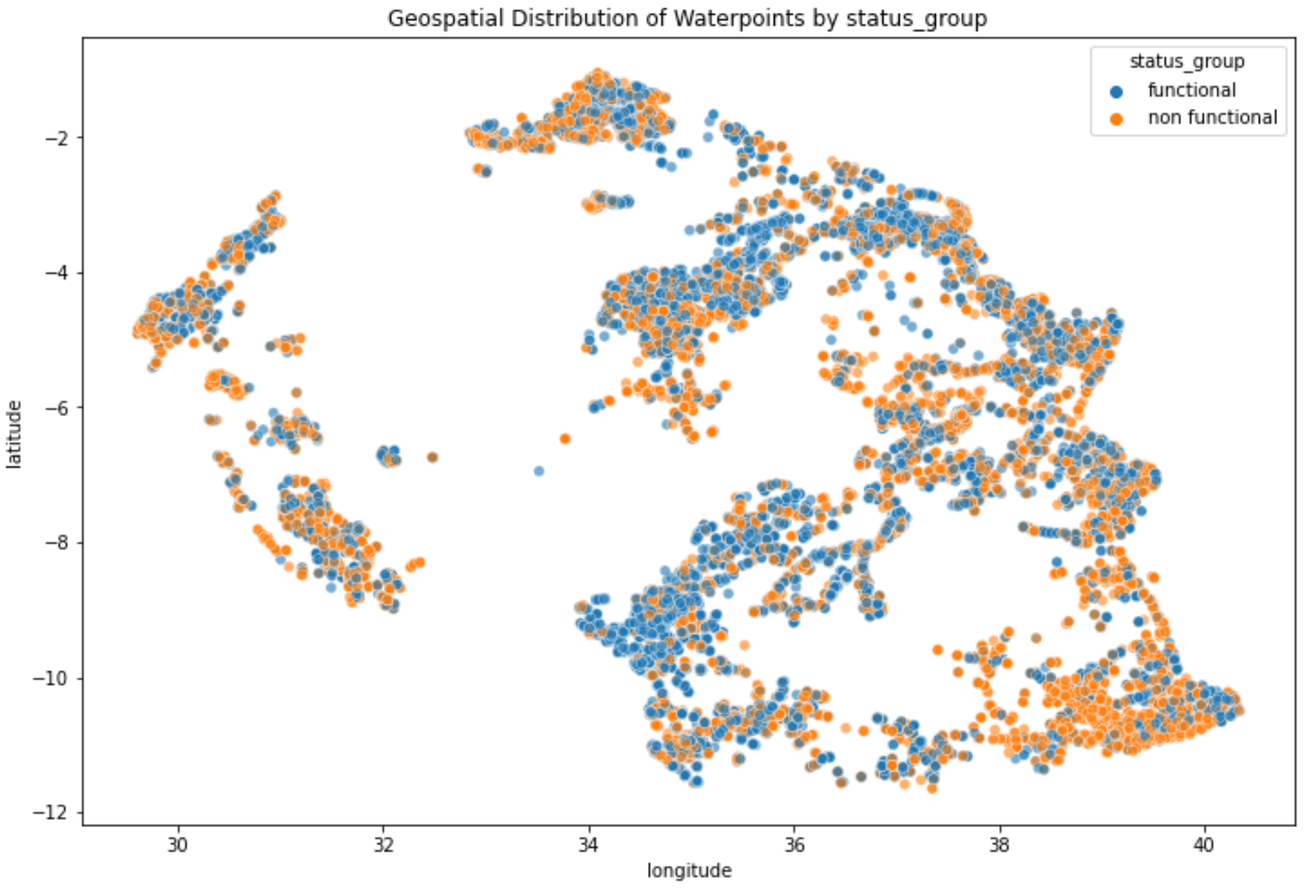
Analysis

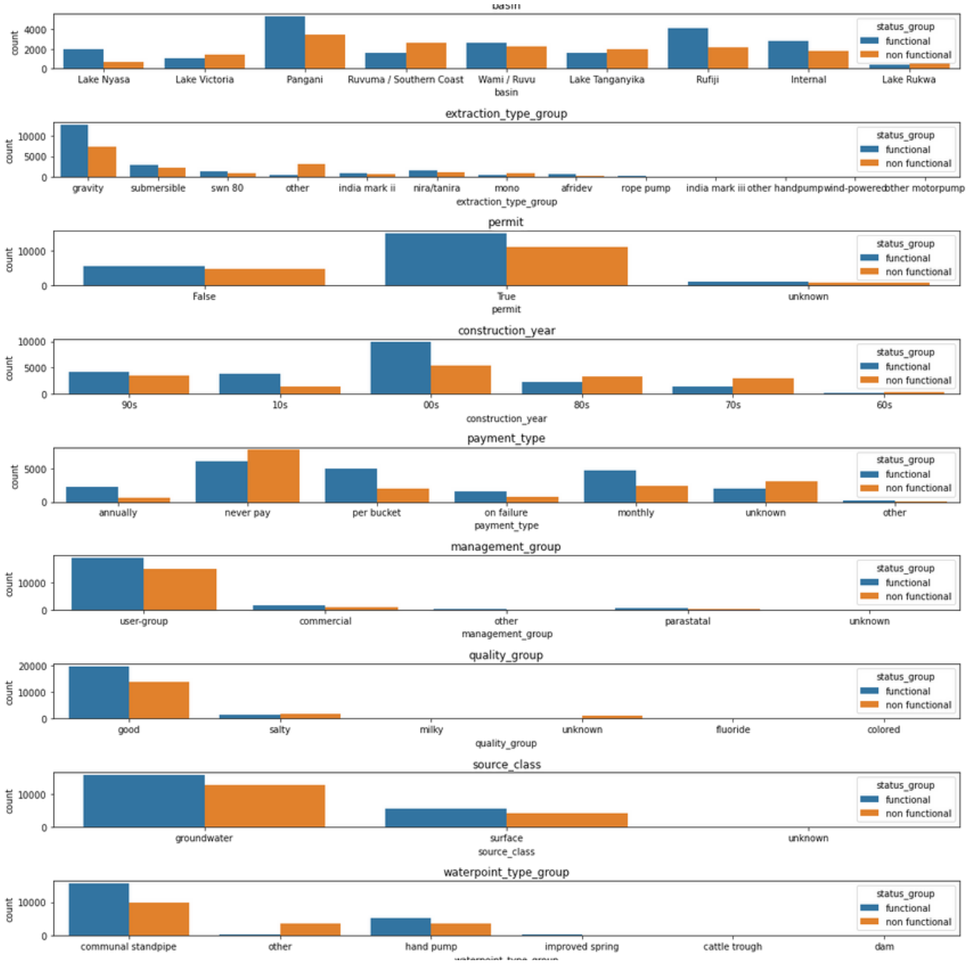
Geospatial analysis using latitud and longitude helps visualize the geographic distribution of water points in Tanzania.verview of your company here.

Histograms provide a concise overview of data spread and reveal underlying patterns.

Count plots offer insights into categorical data distribution in key columns, aiding feature selection and class imbalance identification.

Back to Agenda


Geospatial Distribution of Waterpoints by status_group
status_group
functional
non functional
latitude
longitude
−2
−4
−6
−8
−10
−12
30
32
34
36
38
40


basin
count
4000
2000
0
Lake Nyasa
Lake Victoria
Pangani
Ruvuma / Southern Coast
Wami / Ruvu
Lake Tanganyika
Rufiji
Internal
Lake Rukwa
basin
status_group
functional
non functional
extraction_type_group
count
10000
5000
0
gravity
submersible
swn 80
other
india mark ii
nira/tanira
mono
afridev
rope pump
india mark iii
other handpump
wind-powered
other motorpump
extraction_type_group
status_group
functional
non functional
permit
count
10000
0
False
True
unknown
permit
status_group
functional
non functional
construction_year
count
10000
5000
0
90s
10s
00s
80s
70s
60s
construction_year
status_group
functional
non functional
payment_type
count
5000
0
annually
never pay
per bucket
on failure
monthly
unknown
other
payment_type
status_group
functional
non functional
management_group
count
10000
0
user-group
commercial
other
parastatal
unknown
management_group
status_group
functional
non functional
quality_group
count
20000
10000
0
good
salty
milky
unknown
fluoride
colored
quality_group
status_group
functional
non functional
source_class
count
10000
0
groundwater
surface
unknown
source_class
status_group
functional
non functional
waterpoint_type_group
count
10000
0
communal standpipe
other
hand pump
improved spring
cattle trough
dam
status_group
functional
non functional

Modeling

Logistic Model

Accuracy: 73.74%

SGD Model

Accuracy: 73.90%

XGBoost Model

Accuracy: 81.41%

Tuned Random Forest

Accuracy: 81.45%

Back to Agenda

Conclusion:

- The project successfully addressed water infrastructure challenges in Tanzania through machine learning models, achieving predictive accuracy up to 81.29%
- The developed methodologies provide a foundation for future water infrastructure projects, emphasizing interpretability for effective communication with diverse stakeholders.
- The insights gained enable stakeholders to prioritize maintenance efforts for non-functional wells more efficiently, contributing to sustainable solutions.

Back to Agenda

Recommendations:

- Ensemble Approaches:

Exploring ensemble methods that combine predictions from multip models can enhance robustness and mitigate individual mod limitations. Techniques like stacking or blending could be investigat for improved performance.

- Dynamic Monitoring

Implementing a dynamic monitoring system that continuously updates the model with real-time data ensures adaptability to changing conditions. This would require establishing a reliable data pipeline and periodic model retraining.

Back to Agenda

Next Steps:

- Community Engagement:

Conducting community surveys and engaging with local stakeholders can provide qualitative insights that complement quantitative data. Understanding community perceptions and needs enhances the context of the models' predictions.

- Integration with Decision Support Systems

Integrating the predictive models into decision support systems empowers local authorities and organisations to make timely and informed decisions. This could involve developing user-friendly interfaces for easy access and interpretation.

- Scale to Other Regions:

Extend the project's methodologies to other regions facing similar water infrastructure challenges. Customising the models for diverse contexts broadens the impact and contributes to a more comprehensive understanding of well functionality.

Back to Agenda

Any questions?