Predicting the Condition of Water Wells in Tanzania

Joel Kioko

Problem Statement

- Tanzania faces significant challenges in providing clean and reliable water sources.
- Many water points (wells, boreholes, pumps) are non-functional or in need of repair.
- Limited resources make it difficult to prioritize repairs and maintenance.

Research Objectives

- Analyze the impact of age, technology, and investment on water point failure.
- Assess the impact of socioeconomic and geographical factors.
- Develop a predictive model for water point failure.

Data Sources and Structure

- Data from DrivenData: 3 CSV files (Training set values, Training set labels, Test set values).
- Dataset Size: 59,400 records with 40 features.
- Target Variable: status_group (functional, non-functional, functional needs repair).

Key Insights from EDA

- Target Variable Distribution:
  - Functional: 54.31%
  - Non-functional: 38.42%
- Functional needs repair: 7.27%
- Geographical Distribution: Wells are concentrated in basins like Lake Victoria, Pangani, and Rufiji.
- Water Quality: Majority of wells have "soft" water quality.
- Extraction Types: Gravity and handpumps are the most common.

Data Cleaning and Feature Engineering

- Dropped redundant columns (e.g., id, wpt_name, region).
- Handled missing data:
  Filled missing public_meeting and permit values with False.
- Converted date_recorded to month_recorded.
- Encoded categorical variables using OneHotEncoder.

Model Pipeline and Selection

- Initial Model: Logistic Regression (Baseline accuracy: 80.26%).
- Secondary Model: Decision Tree (Accuracy: 99.85%, but overfitting).
- Final Model: Random Forest Classifier with SMOTE for class imbalance.
- Hyperparameter Tuning: GridSearchCV and RandomizedSearchCV.

Model Evaluation

Random Forest Classifier (RFC):

Accuracy: 79%

Precision (Non-functional): 82%

Recall (Non-functional): 78%

Gradient Boosting Classifier (GBC):

Accuracy: 75%

Precision (Non-functional): 82%

Recall (Non-functional): 62%

Final Model: RFC with tuned hyperparameters.

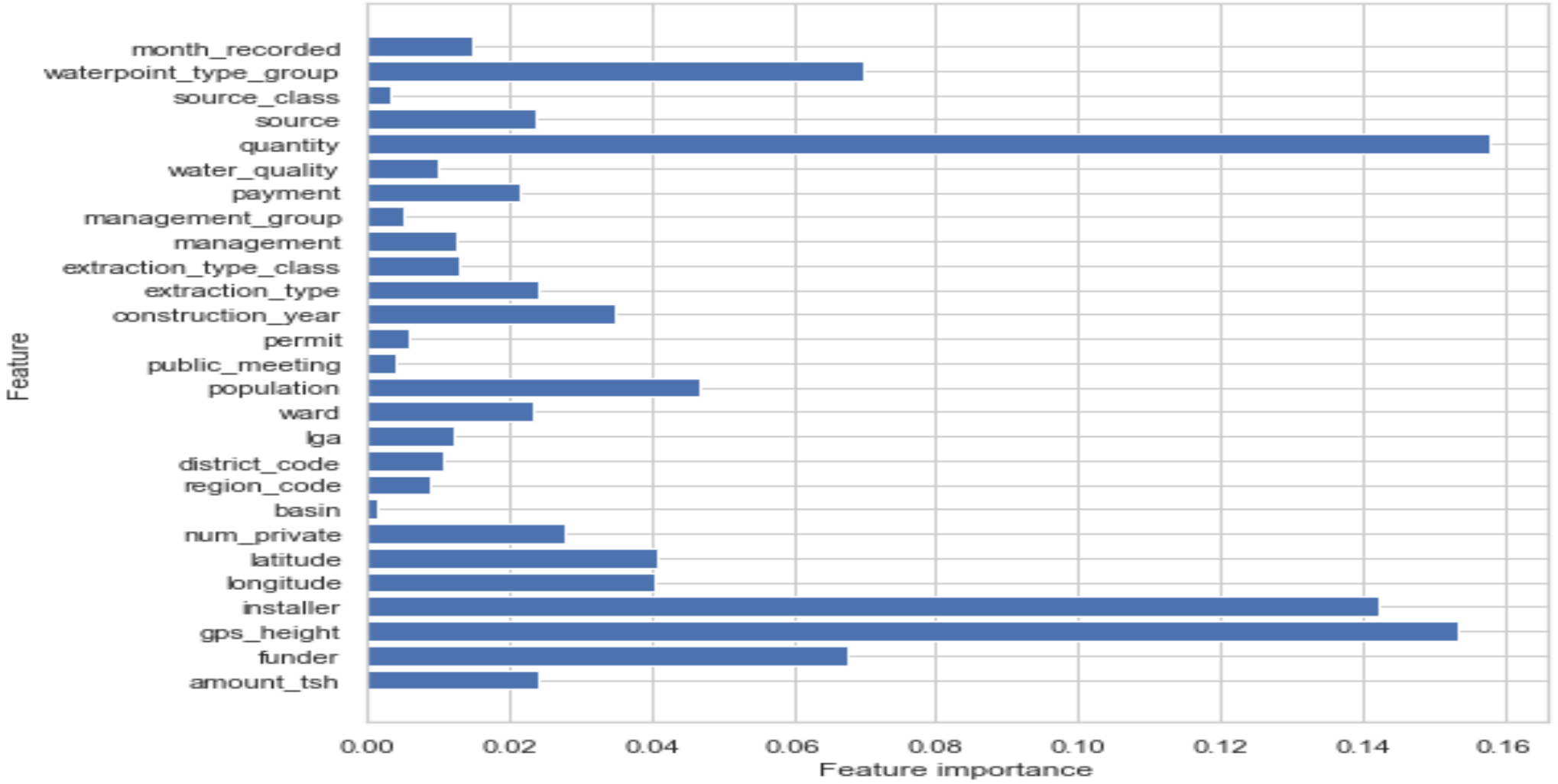
Top Features Influencing Well Functionality

Top Features:

1. Quantity
2. Gps_height
3. Installer
4. Funder
5. Extraction_type_class

Visualization: Horizontal bar chart showing feature importance.

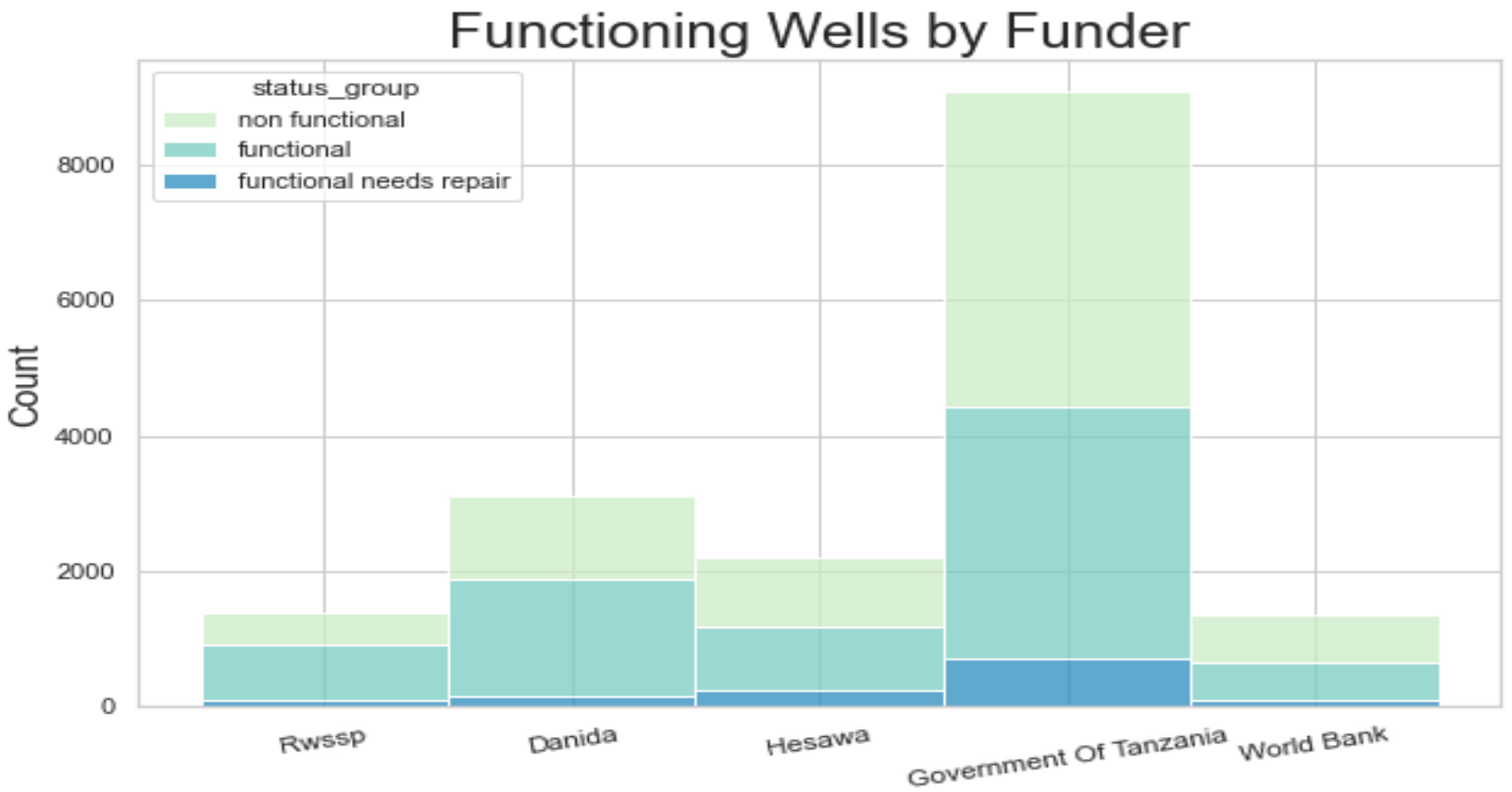
Horizontal bar chart showing feature importance



Impact of Funder and Installer on Well Functionality

- **Top Funders**: Government of Tanzania, Danida, Hesawa, Rwssp, World Bank.
- **Top Installers**: DWE, Government, RWE, Commu, DANIDA.
- **Visualization:** Stacked bar chart showing well status by funder and installer.

Stacked bar chart showing well status by funder



Well Functionality by Basin

- Lake Victoria: 51% functional, 41.59% non-functional.
- Ruvuma / Southern Coast: 37.17% functional, 55.58% non-functional.
- Visualization: Stacked bar chart showing well status by basin.

Well Functionality by Construction Year

- Ruvuma Basin: Wells constructed between 1975-1990 have a higher rate of non-functionality (69.9%).
- Visualization: Histogram showing well status by construction year.

Recommendations

- **Focus on Key Features:** Prioritize wells based on feature importance (e.g., quantity, gps_height).
- **Targeted Interventions:** Collaborate with top-performing funders and installers.
- **Geographical Focus:** Concentrate efforts on basins with high non-functionality rates (e.g., Ruvuma).
- **Historical Data Utilization:** Regularly inspect and maintain older wells.
- **Enhanced Data Collection:** Improve data quality for better predictive accuracy.
- **Community Engagement:** Involve local communities in well maintenance.
-

Conclusion

- The Random Forest Classifier with tuned hyperparameters provides the best predictive performance.
- Key factors influencing well functionality include quantity, gps_height, and installer.
- Targeted interventions and improved data collection can enhance well functionality and water access.

Thank You

Contact Info

joelkioko283@gmail.com