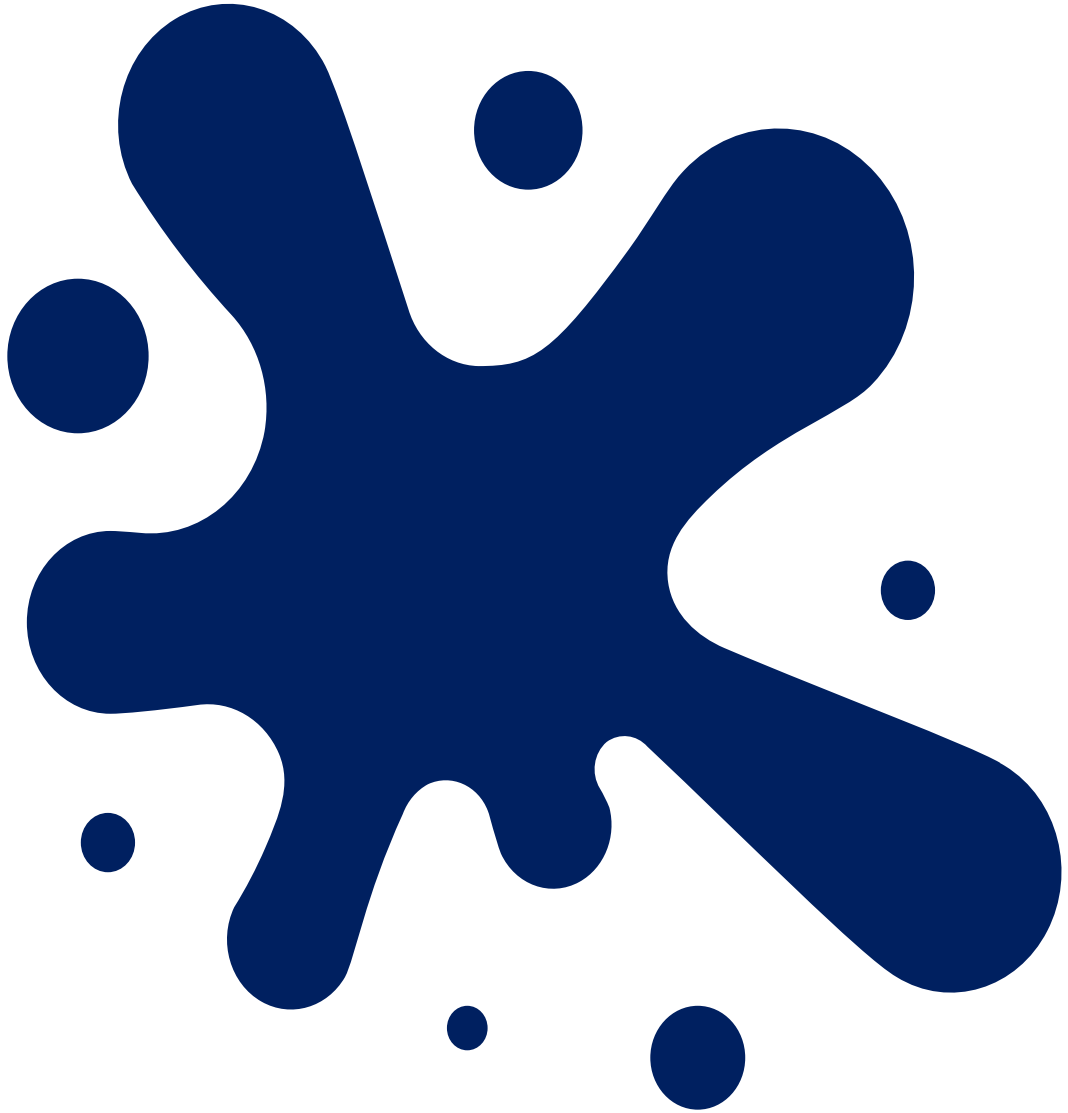


Tanzanian water wells

PREDICTING FUNCTIONALITY

Overview

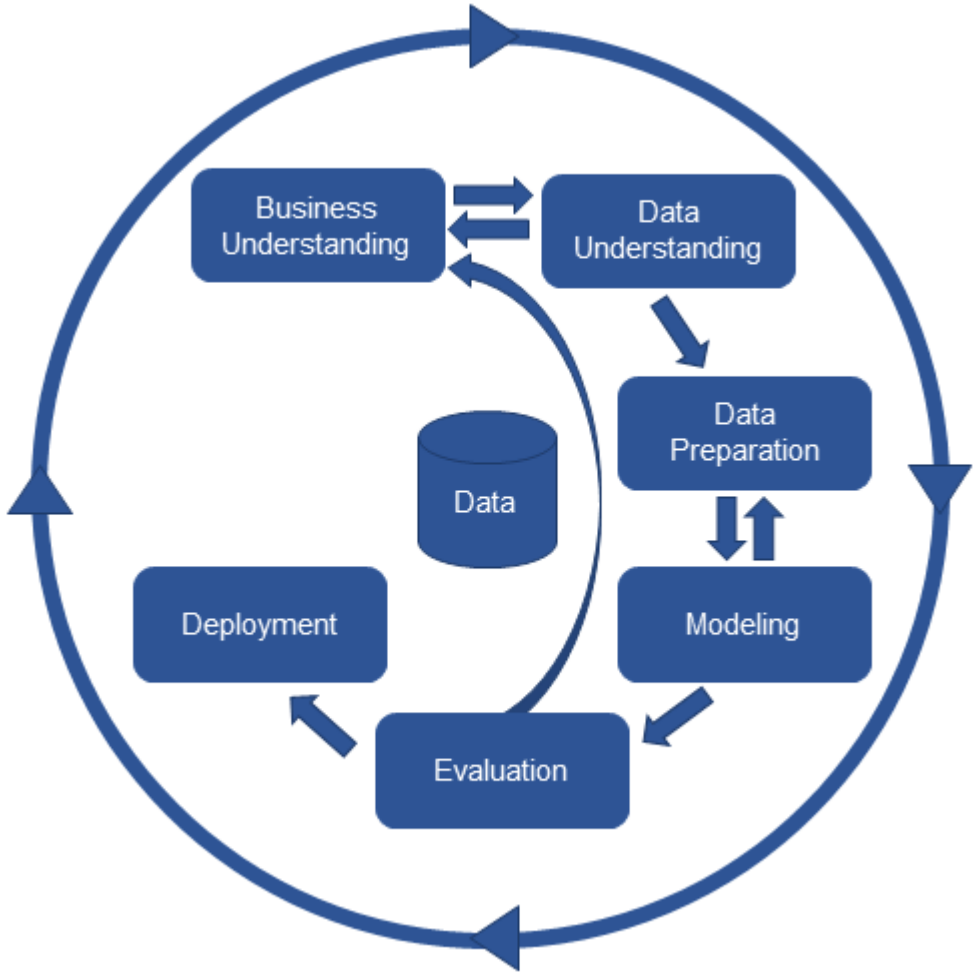
Tanzania is in the midst of a crisis, out of its 65 million population, 55% and 11% of rural and urban population respectively, do not have access to clean water. People living under these circumstances, particularly women and girls, spend a significant amount of time traveling long distances to collect water. This poses significant risks in public health, economic productivity and educational opportunities. Now more than ever access to safe water at home is critical to families in Tanzania

Project objectives

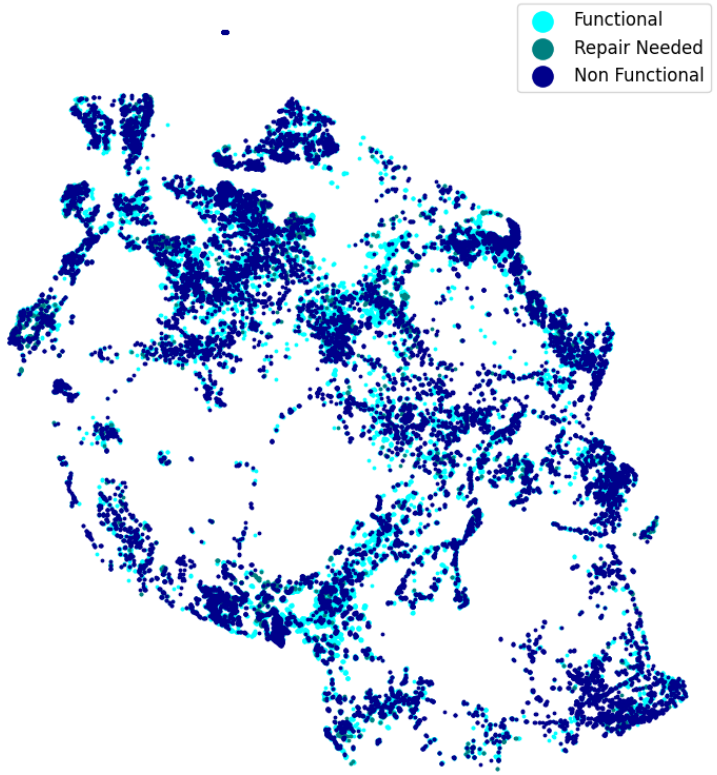
- Understand factors contributing to well performance (geographic, population, funding, etc.).
- Predict well status using machine learning models.
- Optimize model performance for **binary classification** (Functional vs. Non-functional/Needs Repair).
- Provide data-driven recommendations for stakeholders.

Methodology: CRISP-DM

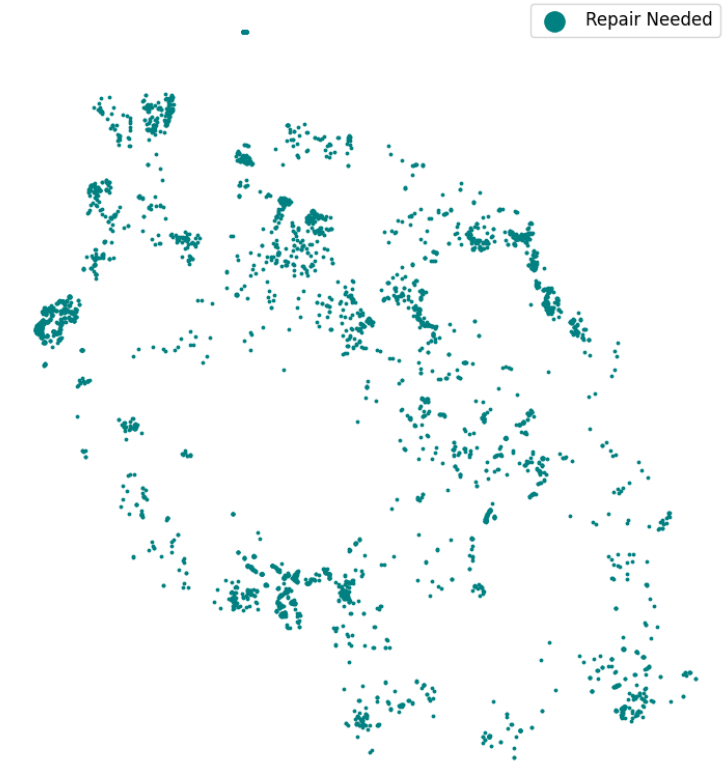
- Data and background information from Taarifa and Tanzania Ministry of Water.
- Initial EDA to understand what the dataset contained
- Data cleaning for modeling
- Initial model creation and results evaluation
- Model benchmarking against the business understanding
- Final model adjustments



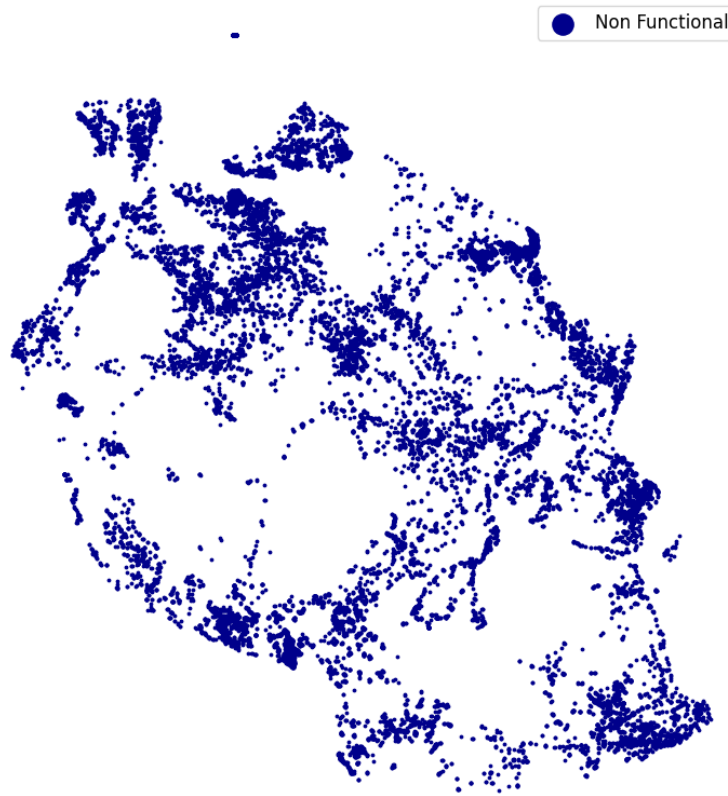
Waterpoint Functional Status



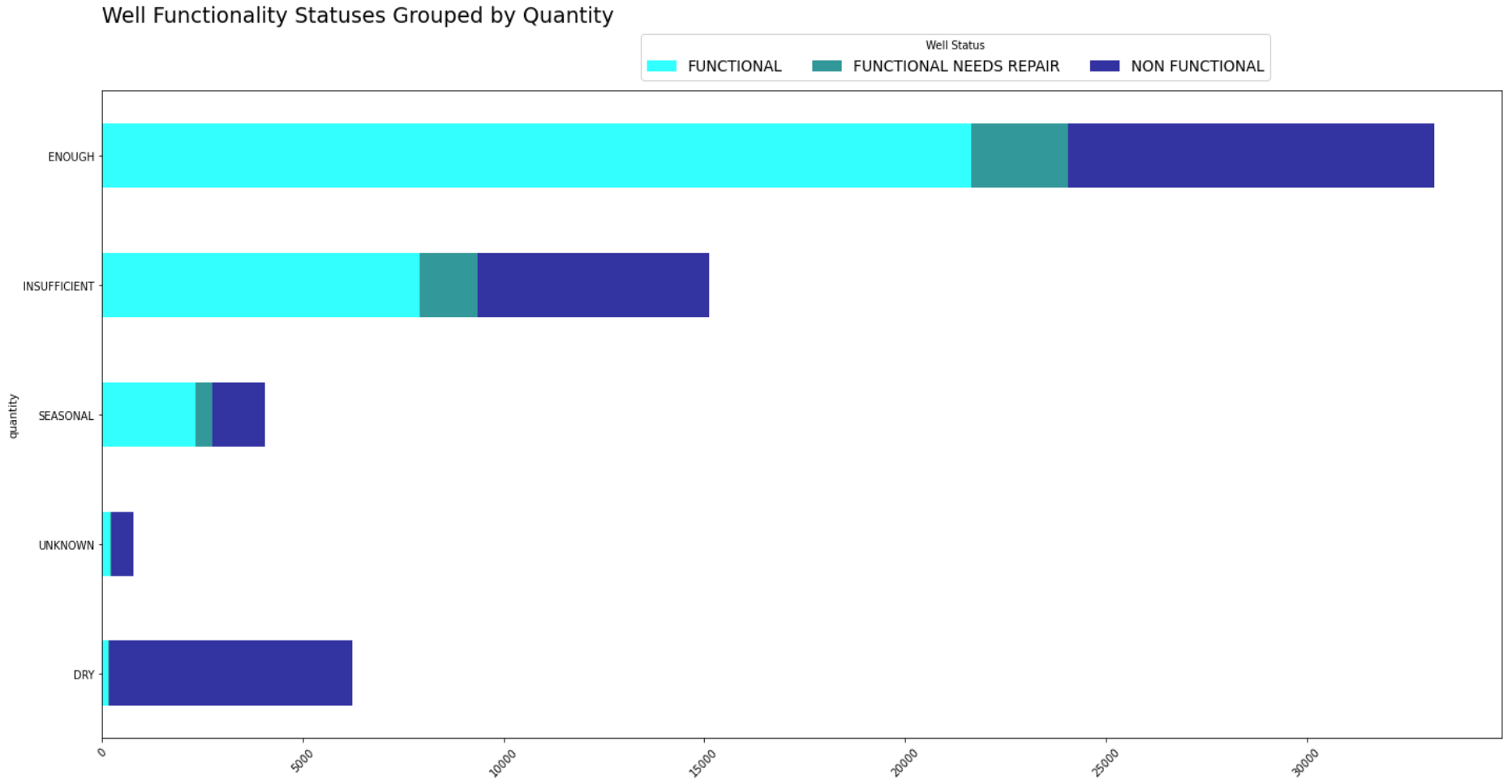
Waterpoint Functional Needs Repair Distribution



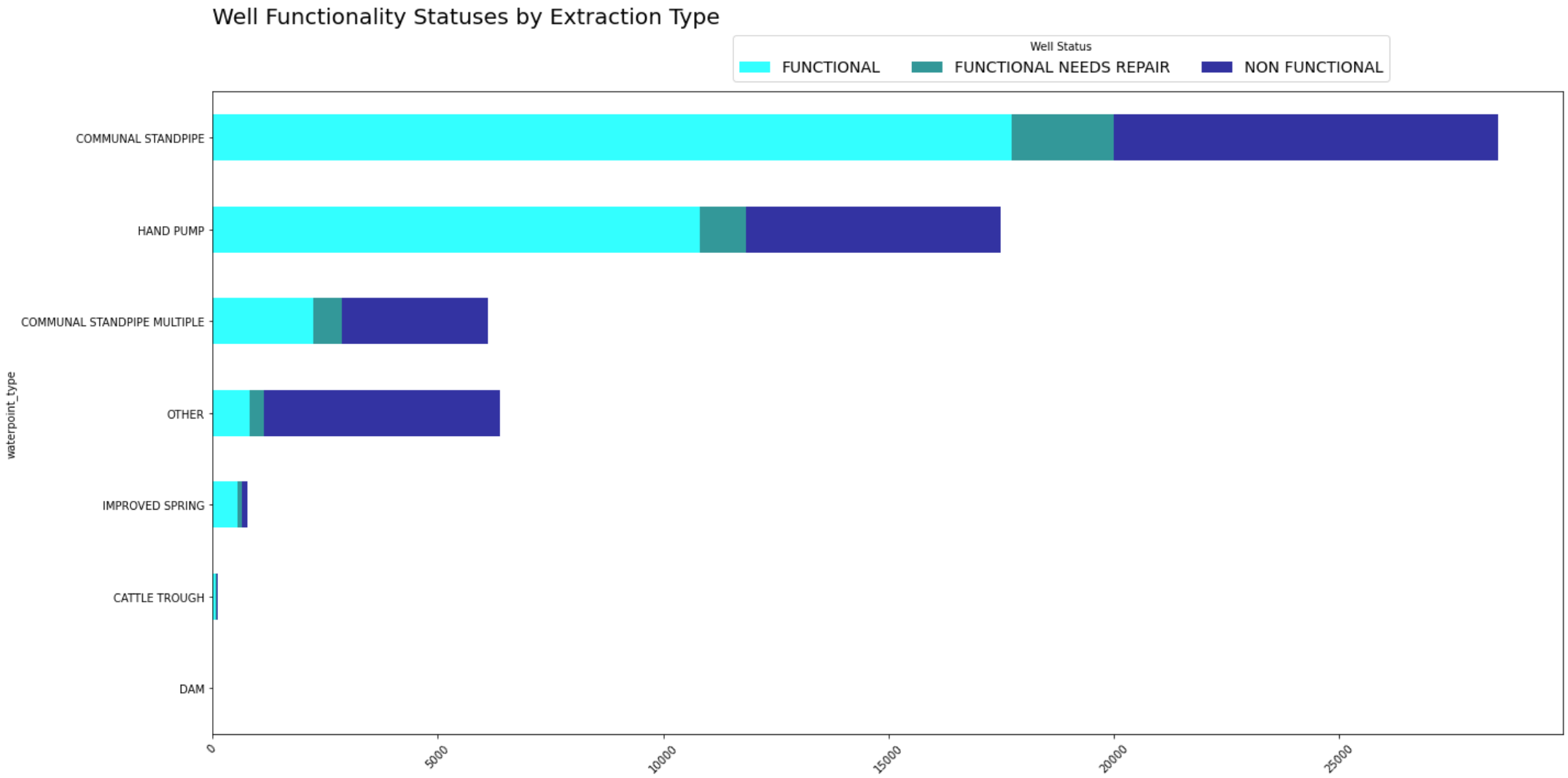
Non-Functional Waterpoint Distribution



The functional but needs repair wells are very underrepresented in our dataset



- Wells with `DRY` quantities will be failed most of the time.
- `UNKNOWN` category also has higher proportion of non-functional wells.



- Wells with waterpoint type `OTHER` are most likely to fail

Modelling & Evaluation

Models tested

- Simple Dummy classifier
- Logistic Regression
- Decision Tree
- Random Forest
- Tuned Random Forest

Models' performance

*Simple Dummy classifier : accuracy of **50%** and recall of **44%***

*Logistic Regression : accuracy of **68%** and recall of **54%***

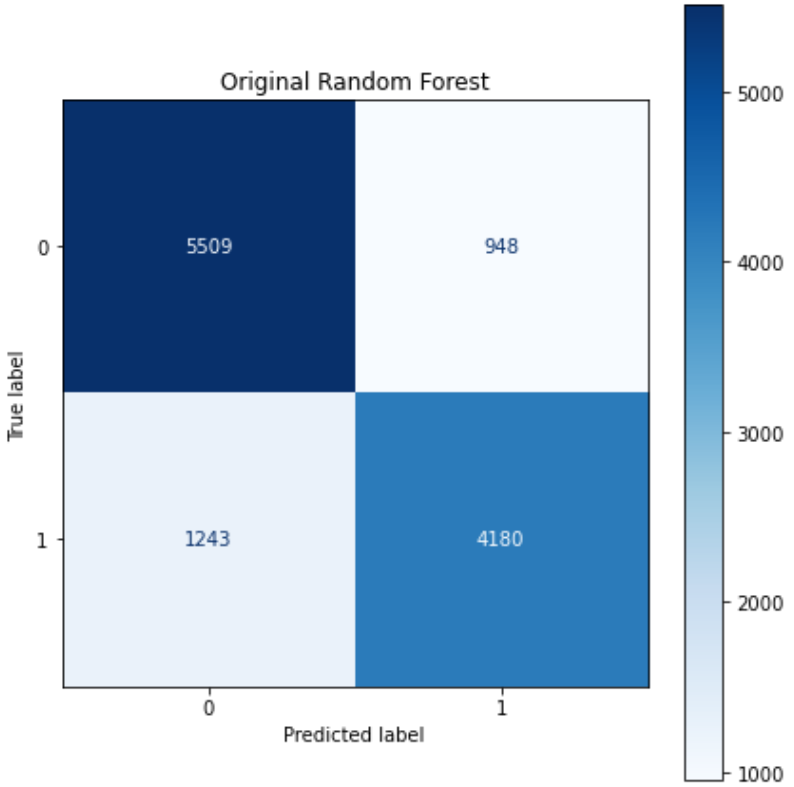
*Decision Tree : accuracy **77%** of and recall of **75%***

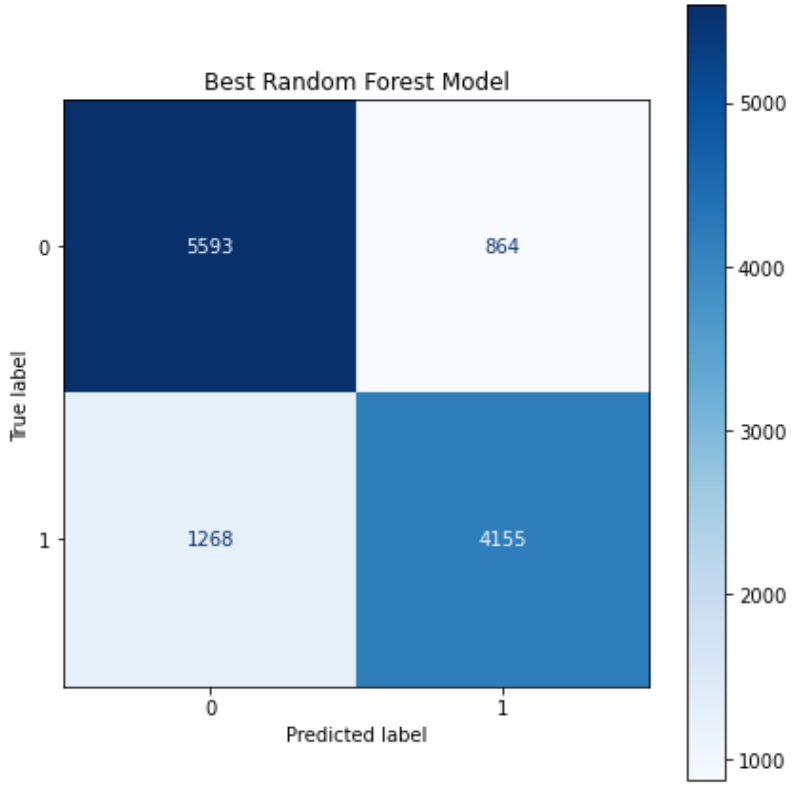
*Random Forest: accuracy of **81.6%** and recall **77%***

*Tuned Random Forest:accuracy **82%** of and recall **77%***

Generally random forest was the best model.

Confusion Matrix-Random forest





- **Accuracy best RF model: 82%**
- **Recall score best RF model: 77%**
- **Precision for 0 (functional) : 82%**
- **Precision for 1 (non-functional/need repair): 83%**

Conclusion and recommendation

- Data quality: Engage with key stakeholders to ensure quality data is collected, to ensure proper handling of outliers or missing values in future with data quality controls put in place.
- Deployment and Monitoring: Develop a robust system for deploying the model into a production environment and monitor its performance over time.
- Focus on class 1 improvement by using advanced techniques such class weighting to improve on dataset balance.
- Further analyze feature importance to actually understand which variable are really driving predictions.
- Explore advanced ensemble methods such as boosting, in order to optimize models' ability to predict minority class.