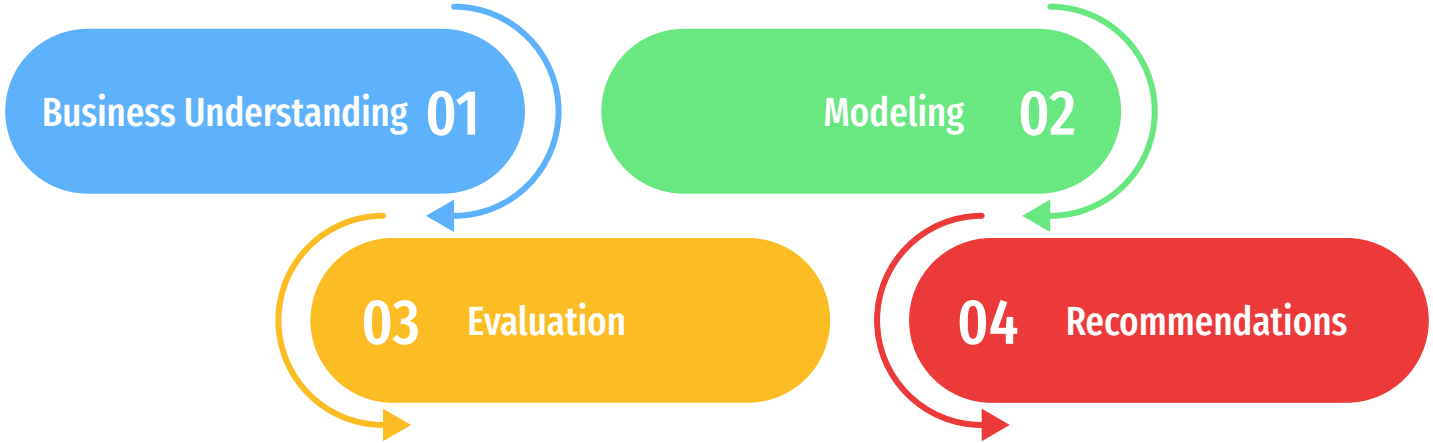
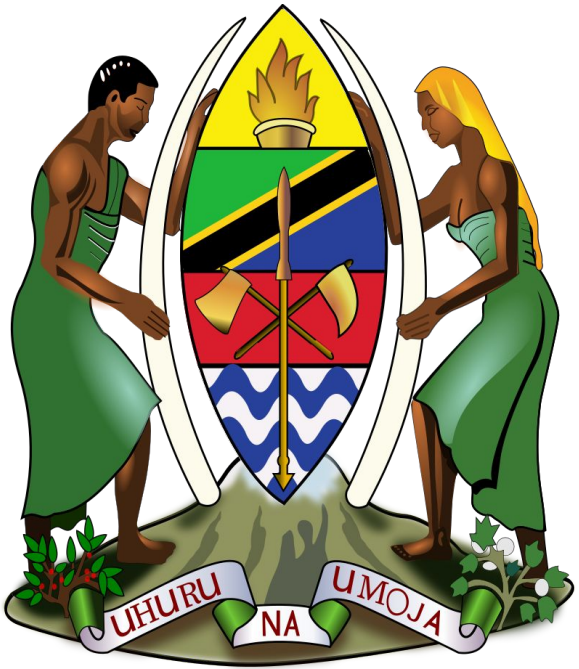
# Tanzania Water Pumps

Eddie, Brian

# Overview


Business Understanding 01
Modeling 02
03 Evaluation
04 Recommendations

# Business Understanding: Our Goal




Using the Tanzania Ministry of Water database on well features, our goal was to create a model that accurately classified whether a pump is functional or not.

Our focus was centered around the specificity score of our model since we wanted to prioritize minimizing the amount of false positives.

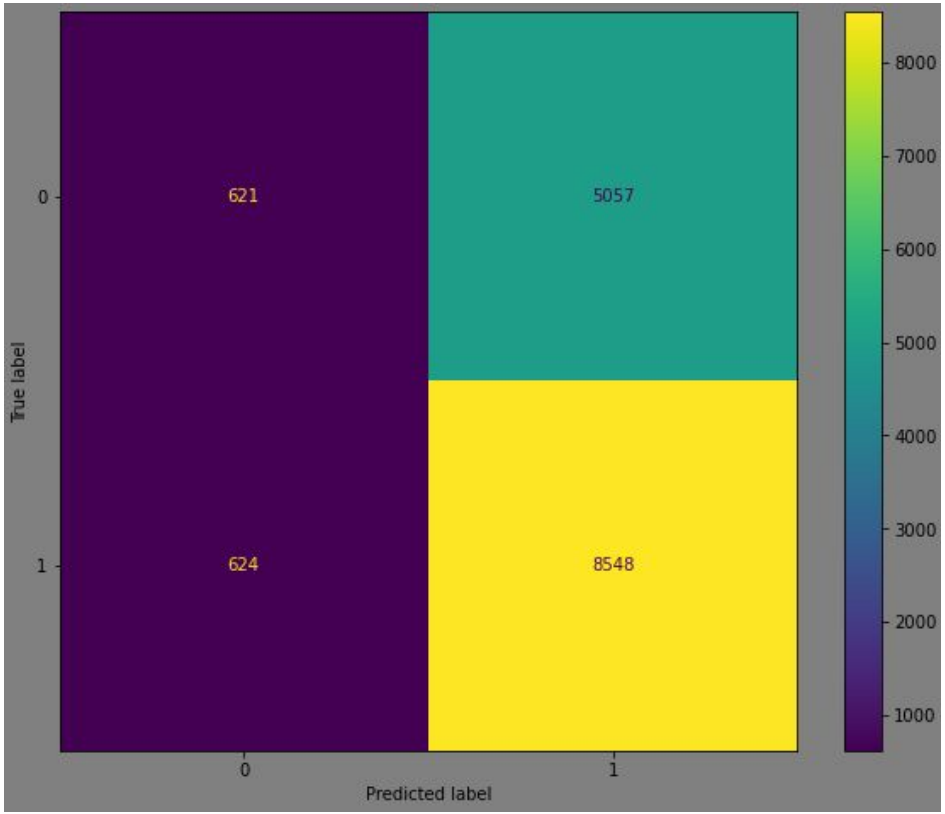
False positives being wells that were deemed functional but in reality were not.

# Numeric Feature Analysis


Small portion of total numeric features

- No combination of factors create clear separation of target variables

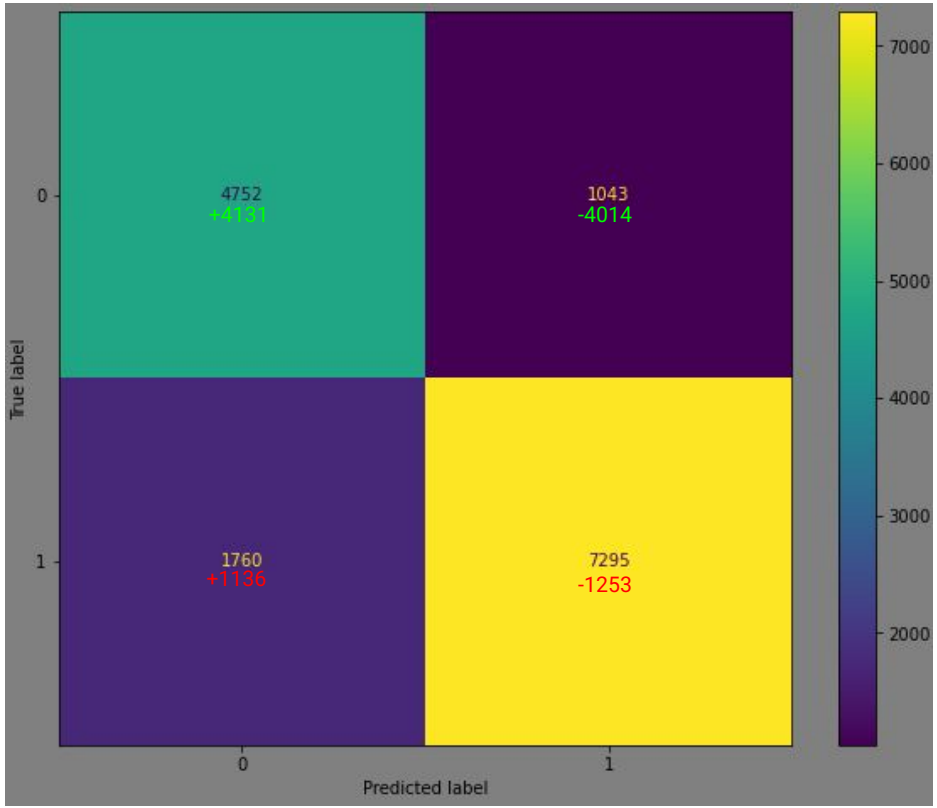
# Baseline Model - Logistic Regression




- Changed labels from ternary to binary
  - Only looked at functional and non-functional wells
- Based only on numerical features(water amount, longitude, etc.)

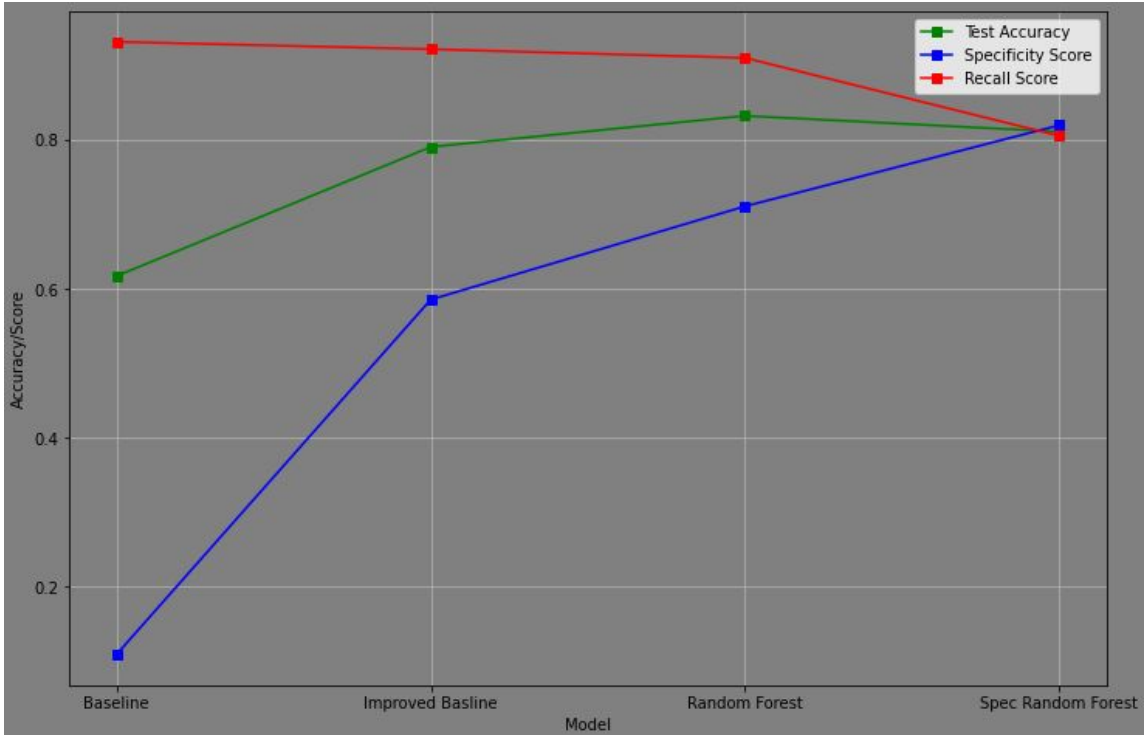
- Accuracy: 62%
- Specificity: 10%
- Recall: 93%

# Second Model -Random Forest Classifier(Specificity)




- Random forest is less prone to overfitting compared to a decision tree
- Include categorical features(region, water quality, etc.)
- Penalize model for predicting a false positive

- Accuracy: 81% (+19%)
- Specificity: 82% (+72%)
- Recall: 80% (-13%)

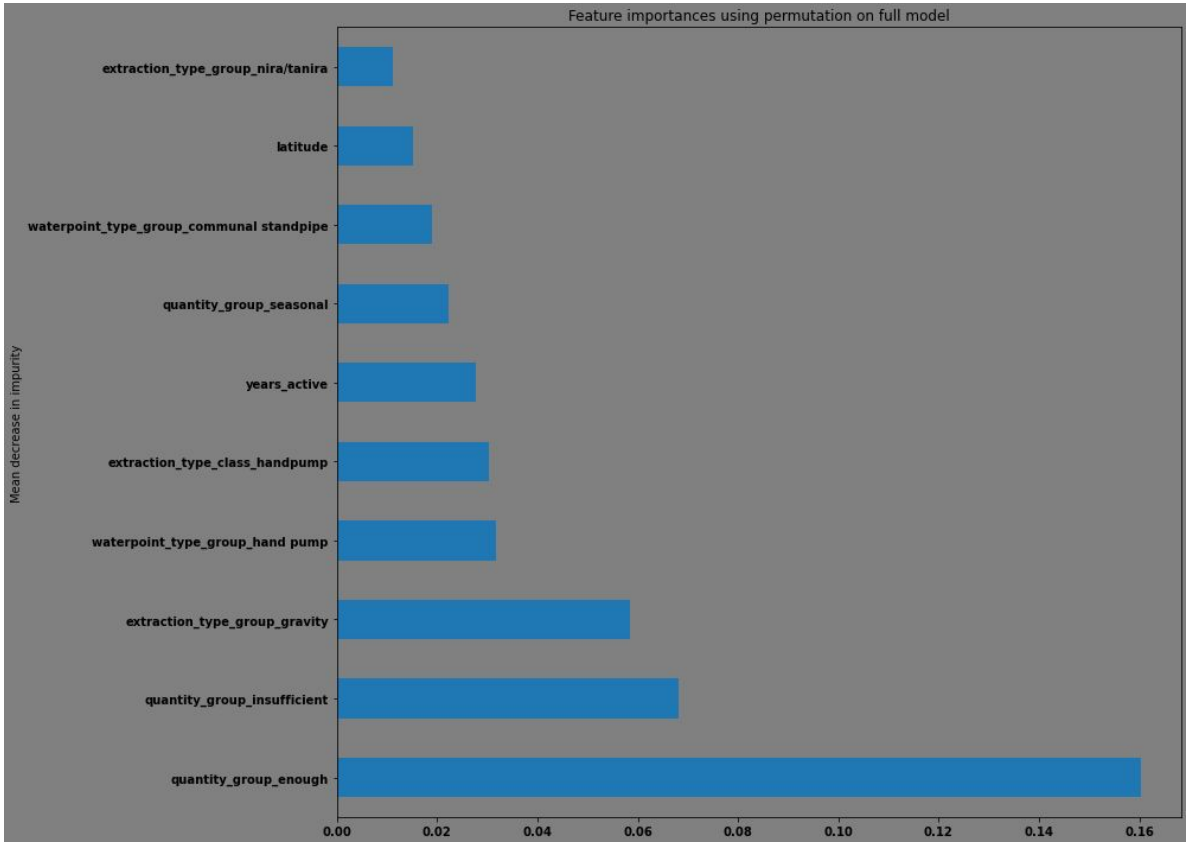



# Accuracy across models

Measures across all models:

- test accuracy,
- Specificity score (Non-functional well prediction)
- Recall score(Functional well prediction)

# Feature Importance - Accuracy




Most important features

- **Quantity_group**: quantity of water
- **Extraction_type**: kind of extraction used at well
- **Years_active**: Years since construction of well to inspection
- **Waterpoint_type**: Type of well(handpump, standpipe, etc.)
- **Latitude**: geographic position

# Next Steps

## Reduce Cardinality

Manually inspect high cardinality features (1000+) and reduce them to be usable in the model

## Focus on Recall

Rebuild model to focus on recall as opposed to specificity since false negatives still require time and resources to check

# Thank you

Questions?