PHASE 3 PRESENTATION

Subject: Tanzanian Water Wells

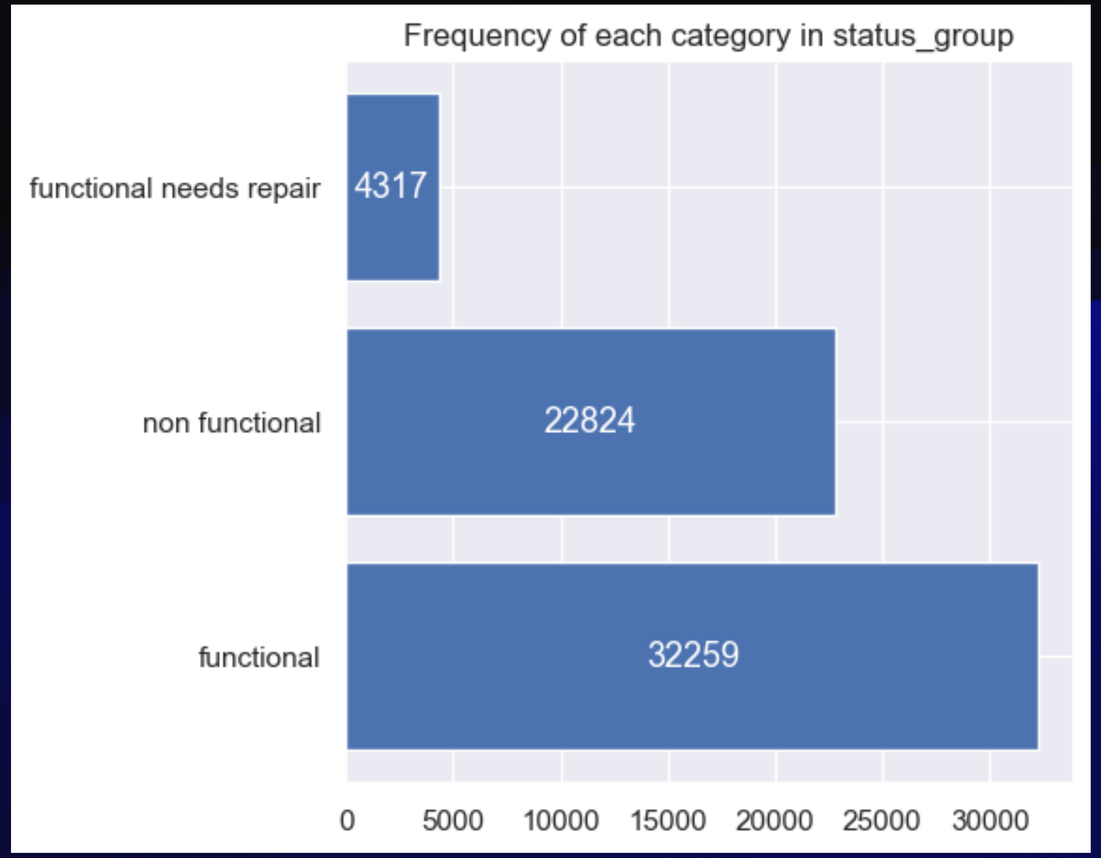
Angelo M. Turri

Data

- Comes from an online competition on www.drivendata.org
- ~59,000 records
- Each record in the data is a single Tanzanian water well
- 40 features
  - 10 numeric features
  - 30 categorical features
- A single target variable

Target Variable – Water well status

- THREE CATEGORIES
  - Functional
  - Non functional
  - Functional needs repair



Stakeholder

- Charity organization with limited funds
  - Their goal is to fix as many water wells as possible in as little time as possible
  - Out of all the water points, the “functional needs repair” and “non functional” wells are the ones that require attention
  - Non functional wells require significantly more resources to fix than functional needs repair wells
  - They need us to predict all three categories with maximum accuracy, so they can decide the amount of resources to send to each water well

Data Preprocessing

- Too many features
- Several variables aren’t suitable for our models
  - Do not correlate with target variable (e.g., id column)
  - Cause collinearity (e.g., “payment” and “payment_type”)
  - Differ in their categories from dataset to dataset

After data pruning

- 18 features – 9 numeric, 9 categorical
  - 2 engineered numeric features
- All categorical variables were one-hot encoded
- All numerical variables were scaled

Important numeric features

- Location
  - Longitude, latitude, altitude
- Surrounding population
- Year of construction

Important categorical features

- Location
  - Tanzania was divided into 124 provinces
- Water quality
- Water extraction method
- How it receives payment

Metrics

- F1 score for each status category (balanced average of precision and recall)
  - Precision is how likely you are to be right when you predict a certain category
  - Recall is how many instances of a category you successfully identified
- % of non functional wells incorrectly classified as functional

Estimators used

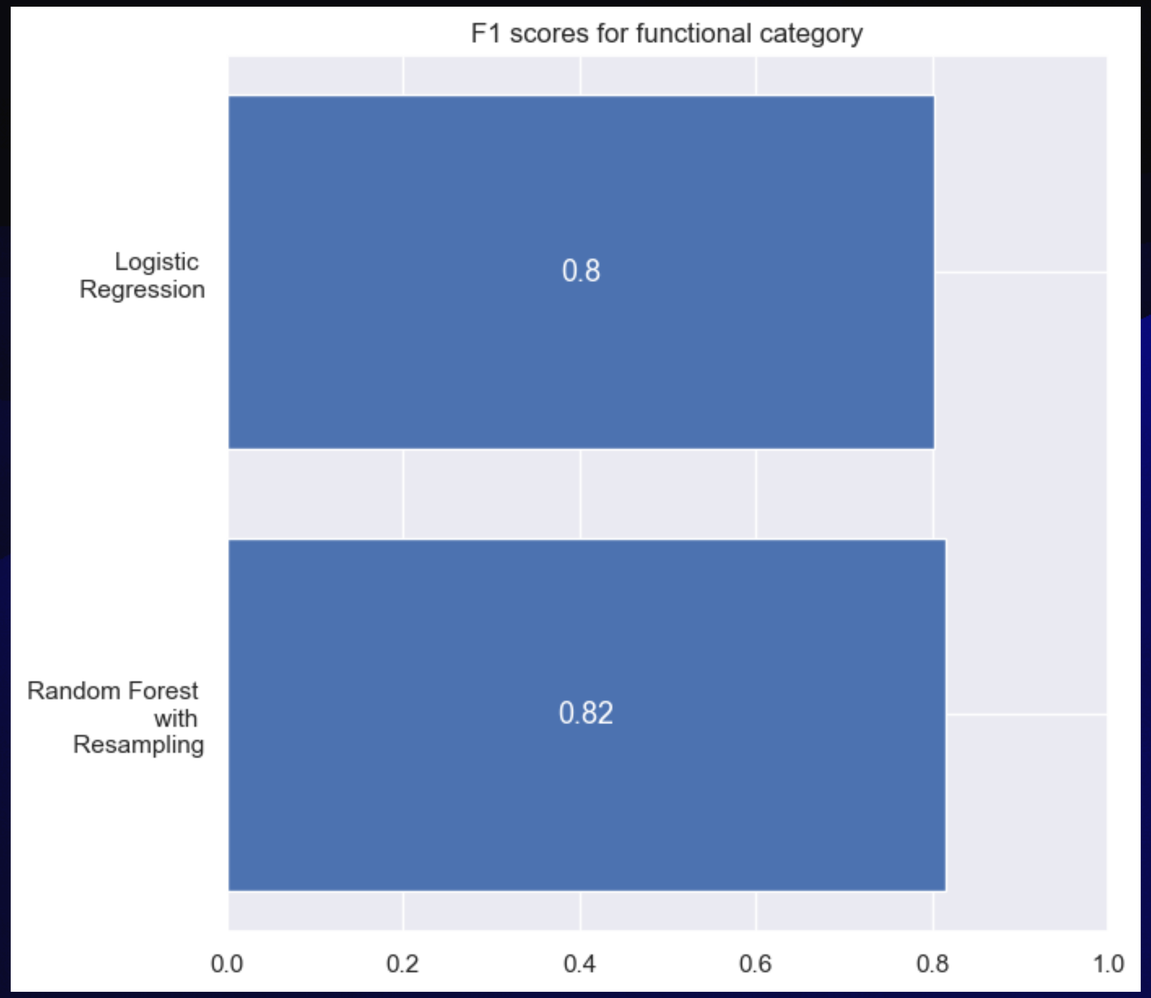
- Iterative modeling approach – started with a very simple model and made improvements to it based on metrics
- Initial model: Logistic Regression, unsatisfactory results
- Best model: Random Forest with 6,860 resampled instances of the “functional needs repair” category

SMOTE explanation

- Creates synthetic data
- Like recycled paper. No new material is used, but a bunch of old material is mixed around and re-used.

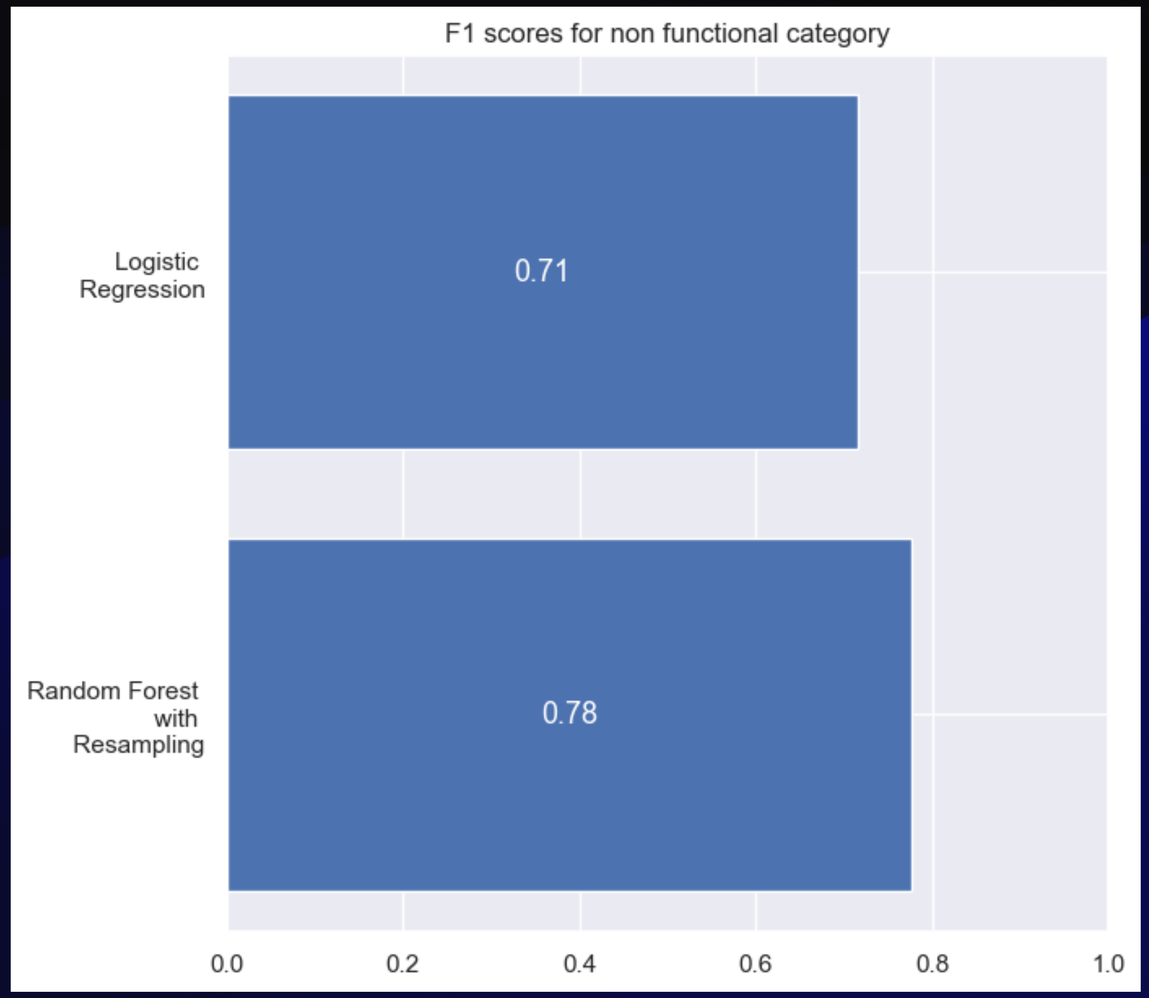
Comparing F1 scores: “Functional” category

- F1 Score improved by 0.02 with Random Forest



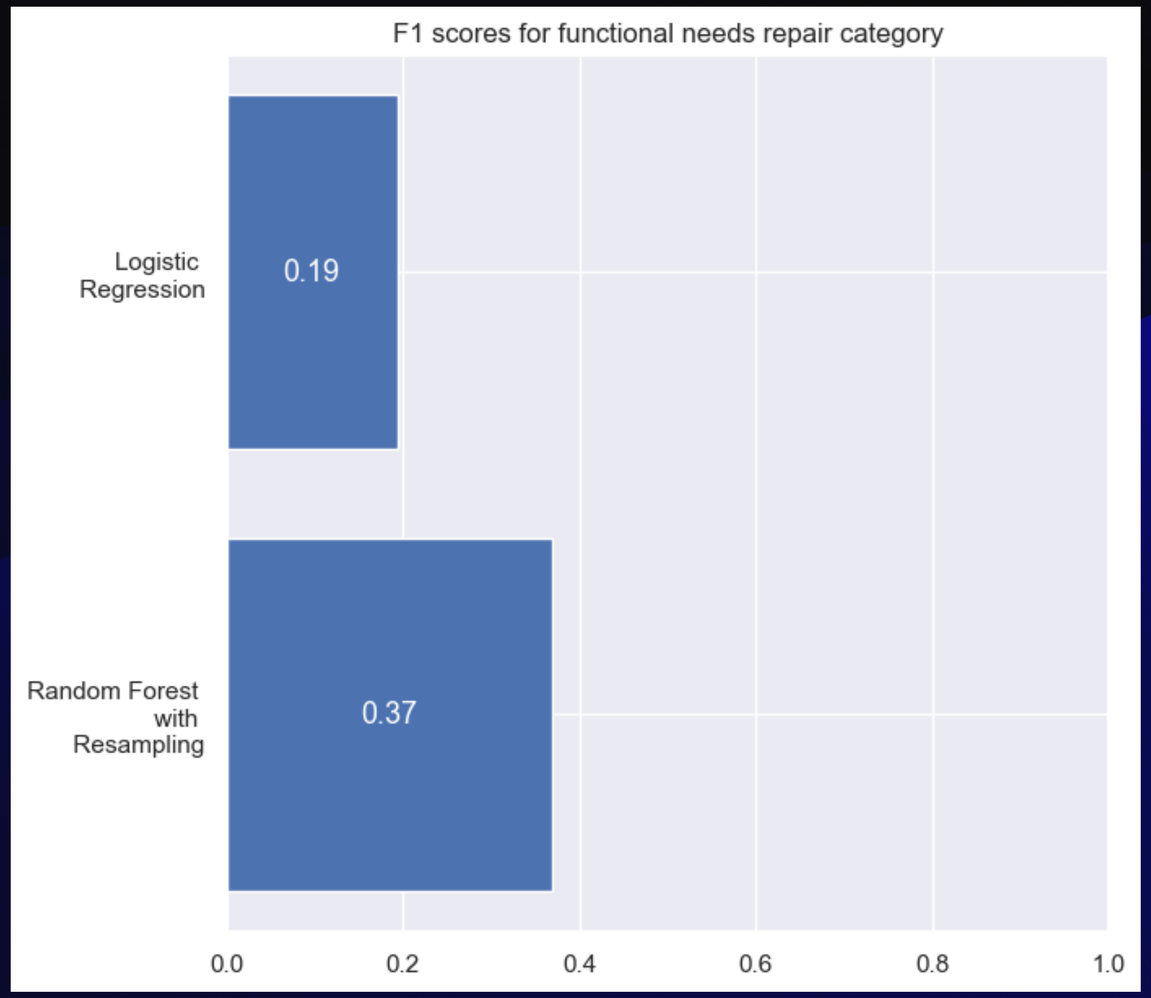
Comparing F1 scores: “Non functional” category

- F1 Score improved by 0.07 with Random Forest



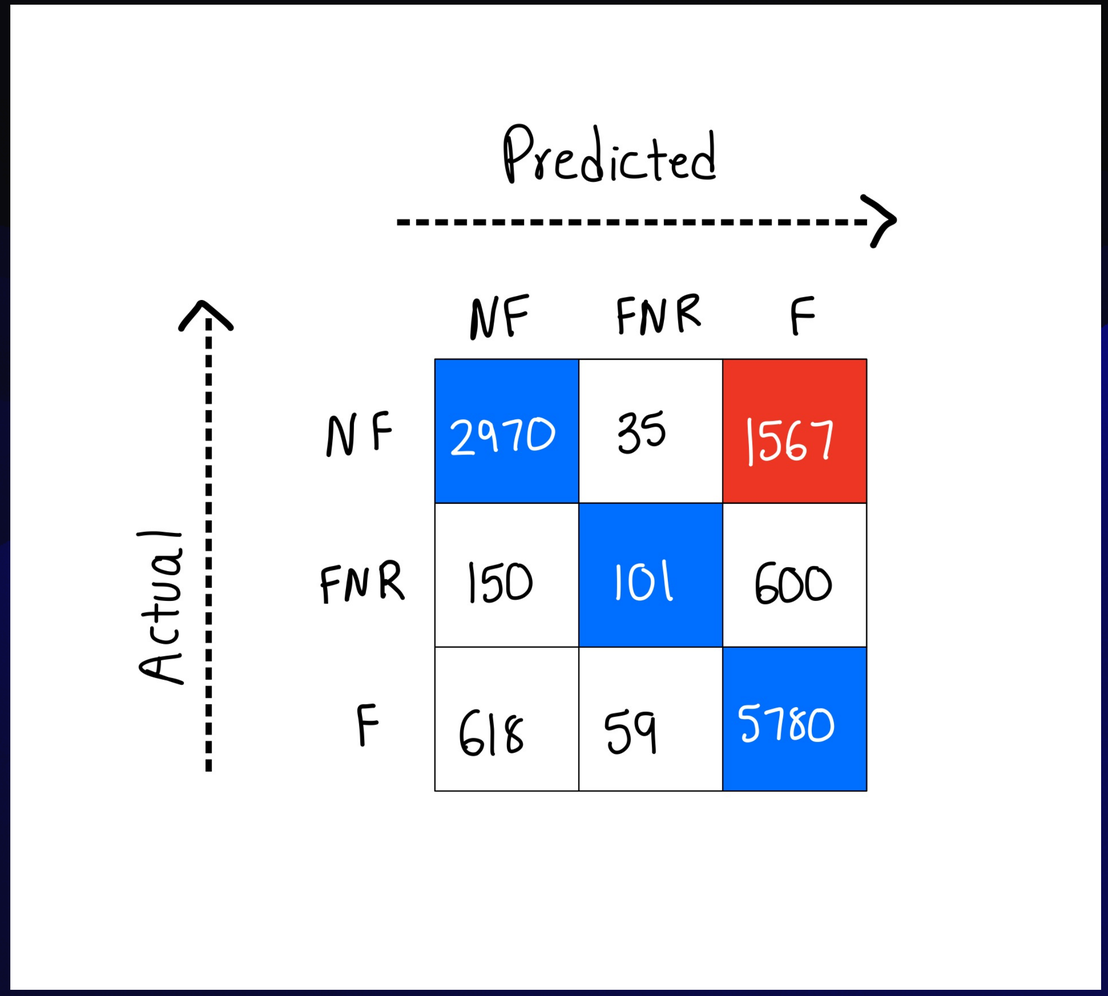
Comparing F1 scores: “Functional needs repair” category

- F1 Score improved by 0.16 with Random Forest
- Highest F1-score improvement across all three categories



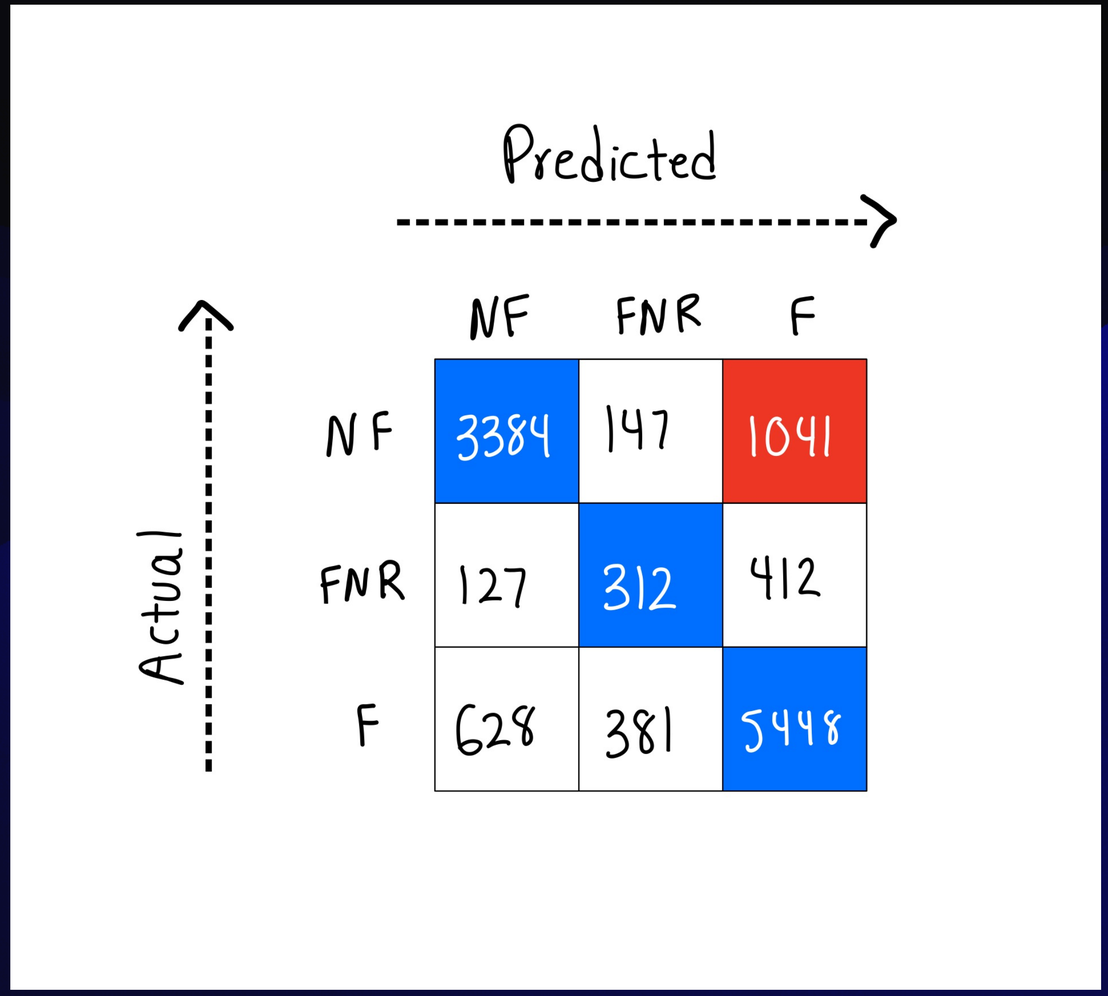
Confusion Matrix: Base Model (Logistic Regression)

- Blue square is best
- Red square: classifying non functional well as functional



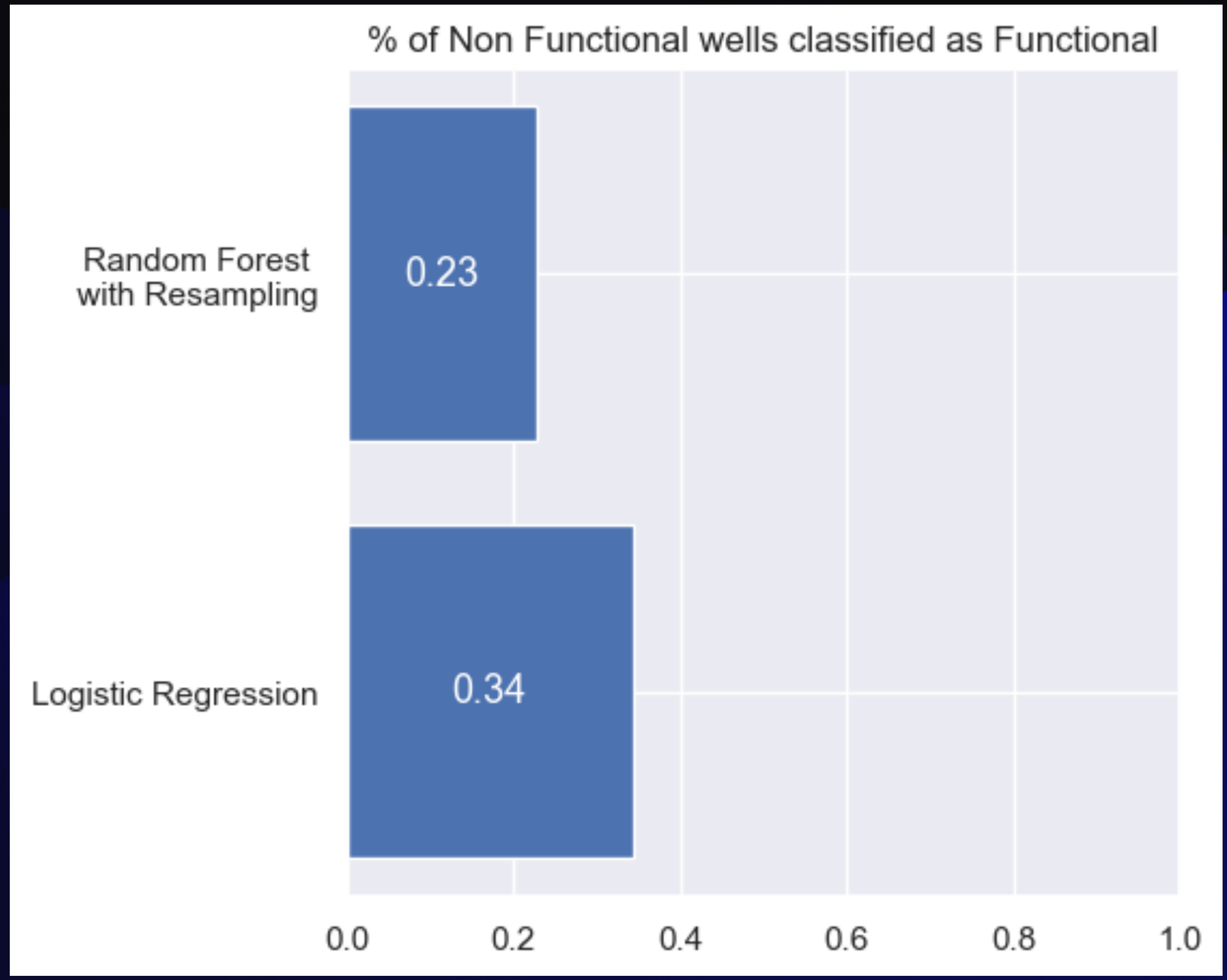
Confusion Matrix: Best Model (Random Forest)

- More grouped on blue diagonal
- Less in red square

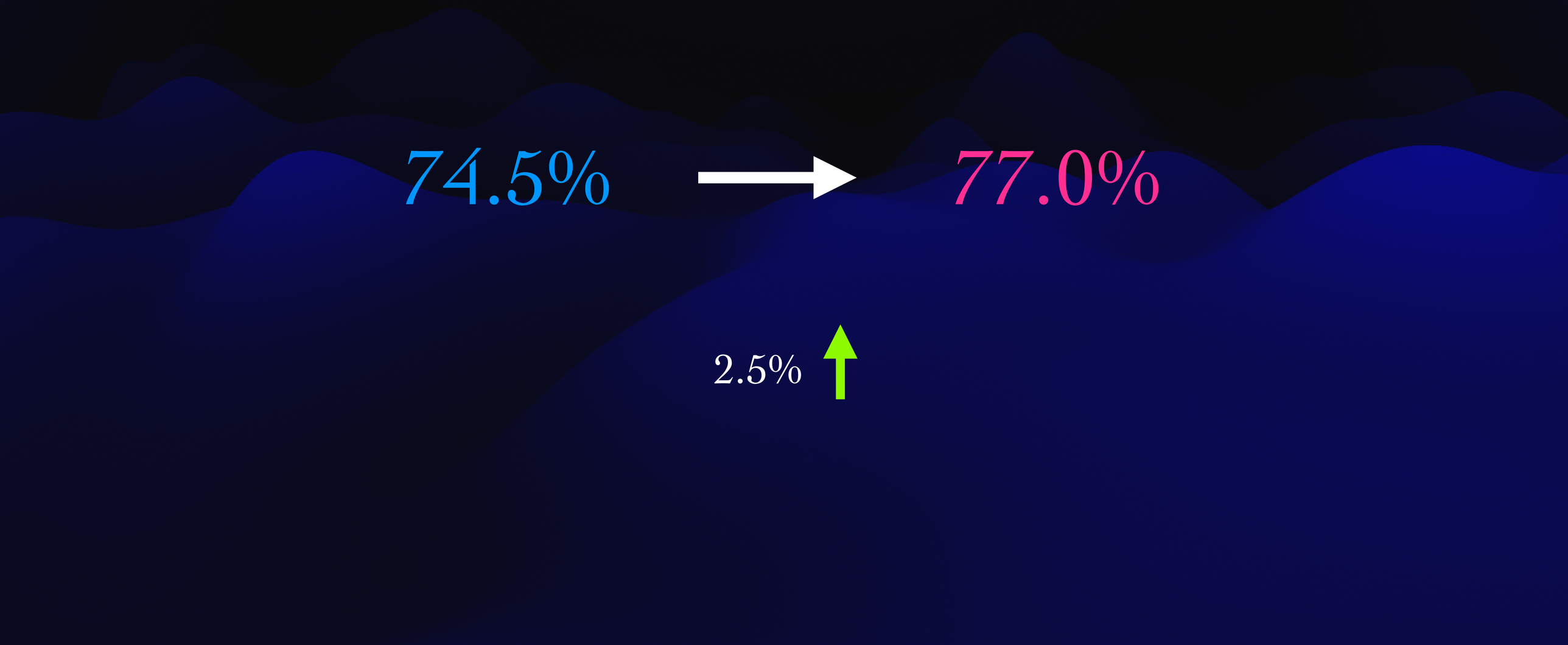


Reducing NF -> F error

- Classifying a non-functional well as functional leaves a community without water
- The percentage of non functional wells mistakenly classified as functional was reduced by 11%

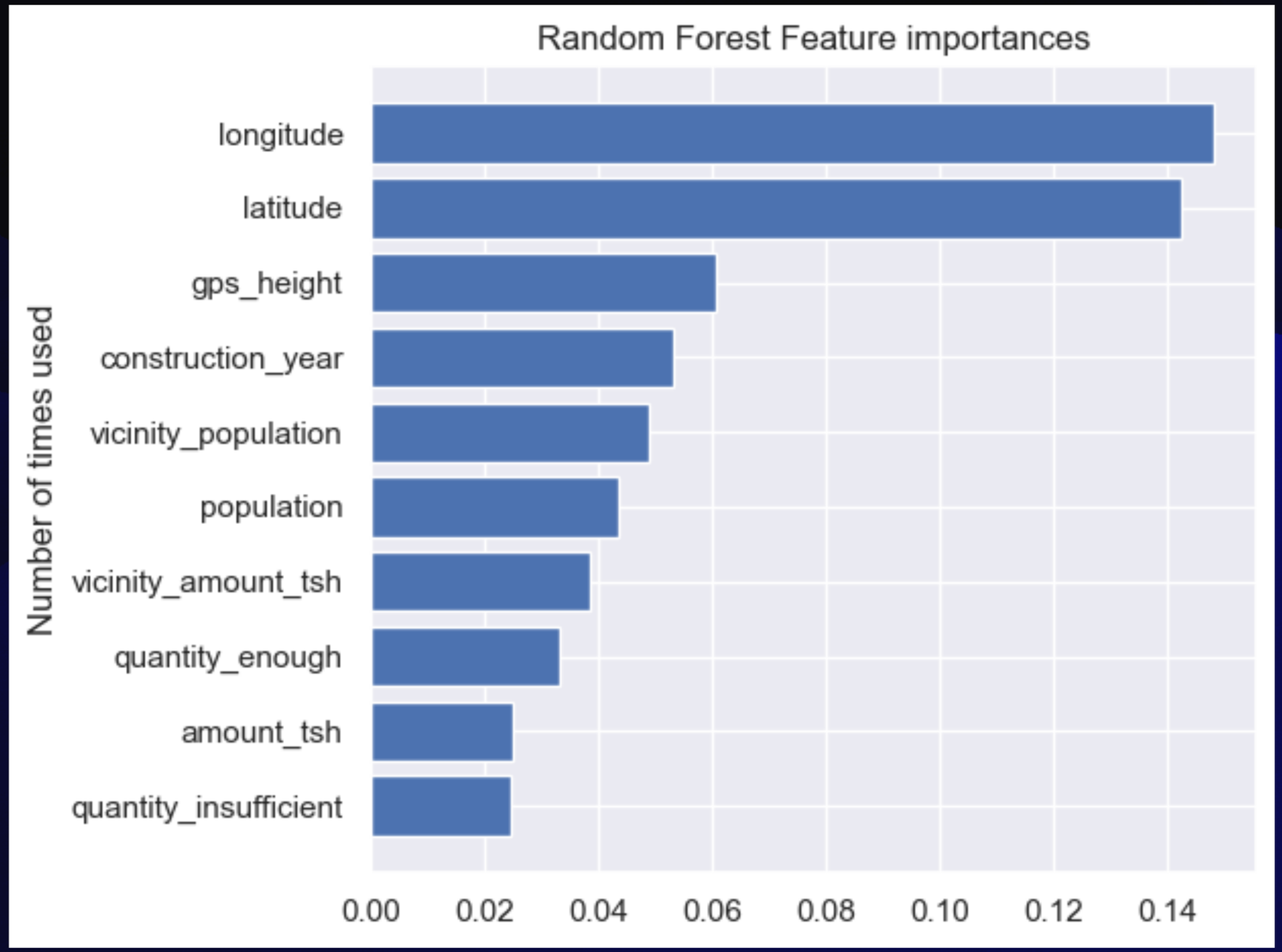


Comparing Overall Accuracy



Feature Importances

- The variables that our model deemed most important
- Location, population, construction year critical

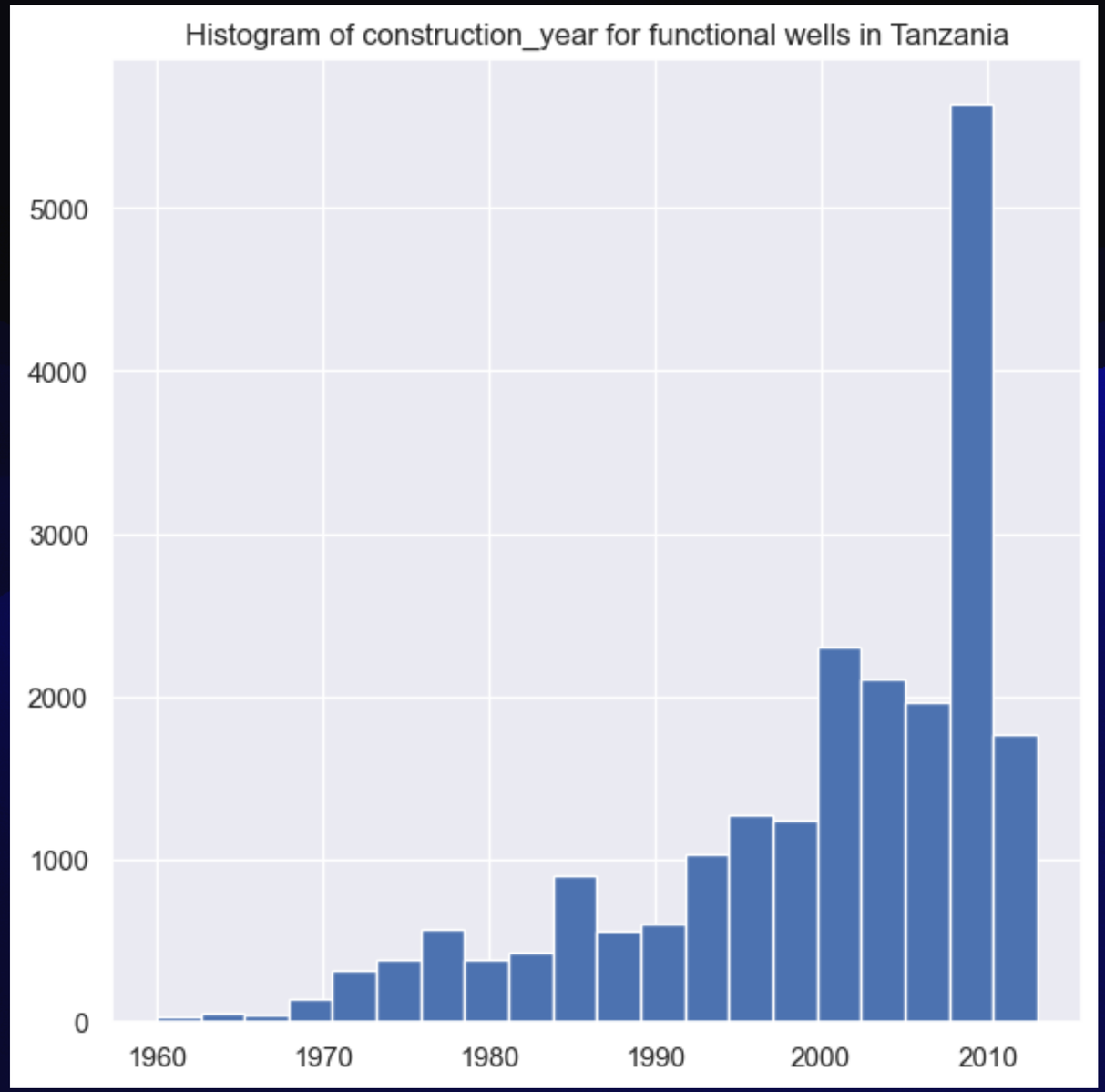


Providing the Requested Predictions

- This is where I would provide the requested predictions if necessary
- If it matters, I submitted them on drivendata.org, which is where the original competition is being hosted, and I placed roughly #4600 out of almost #16,000 participants.

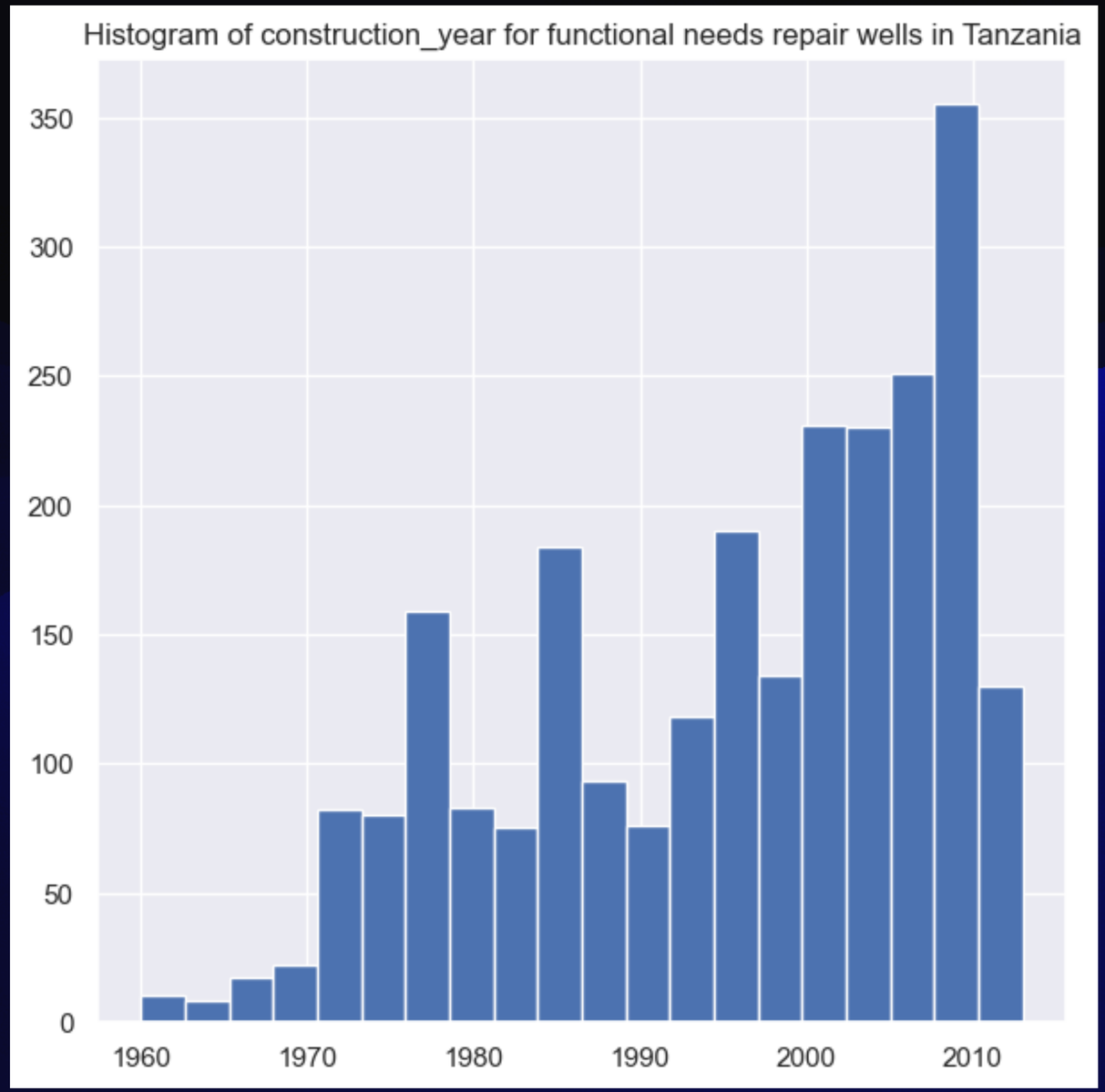
Construction year distribution for “functional” category

- Concentrated mostly on right-hand side
- Most of the wells are newer



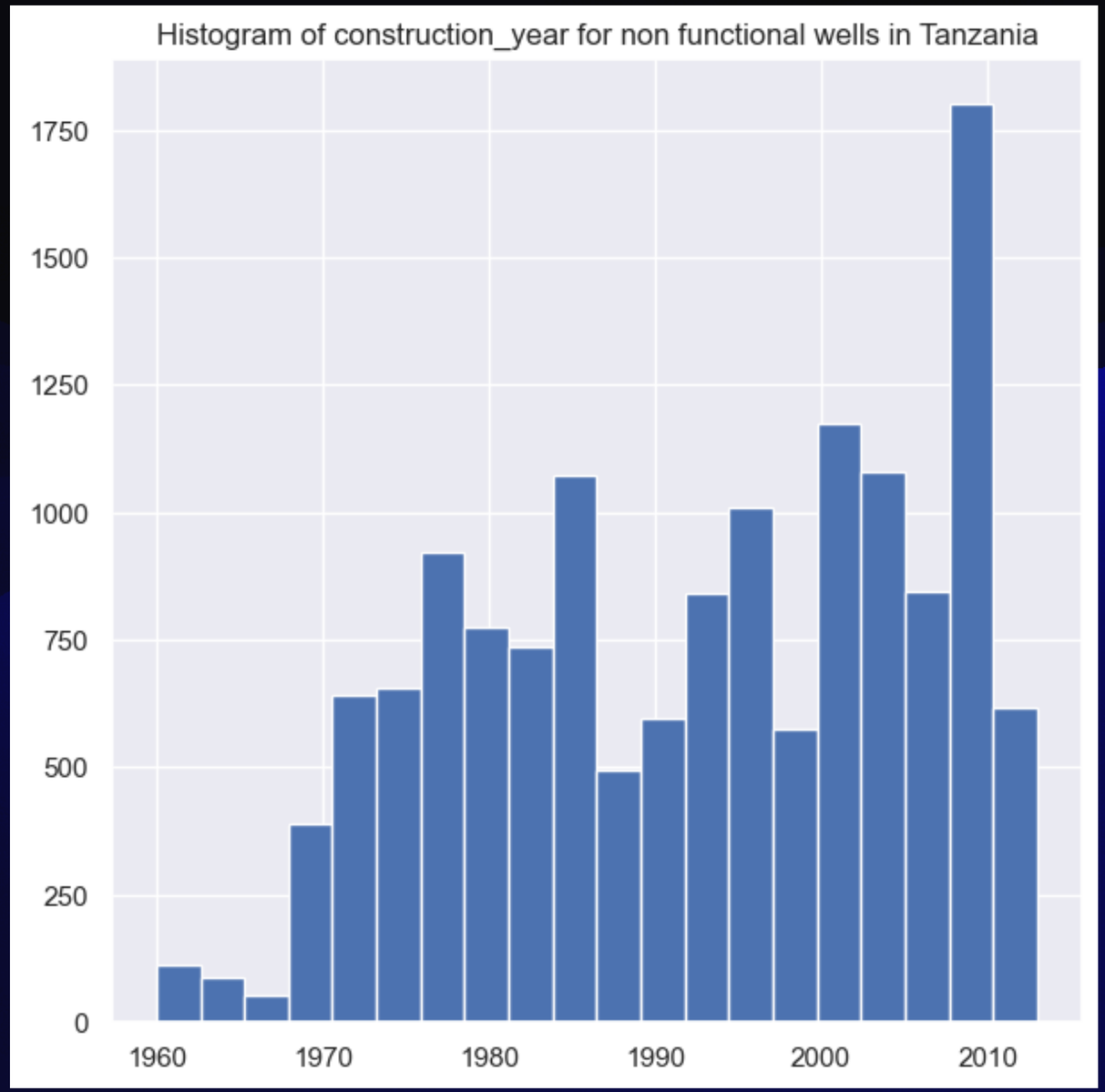
Construction year distribution for “functional needs repair” category

- The distribution shifted to the left
- More of the wells are older



Construction year distribution for “non functional” category

- Shifted to the left yet again
- Even more of the wells are older

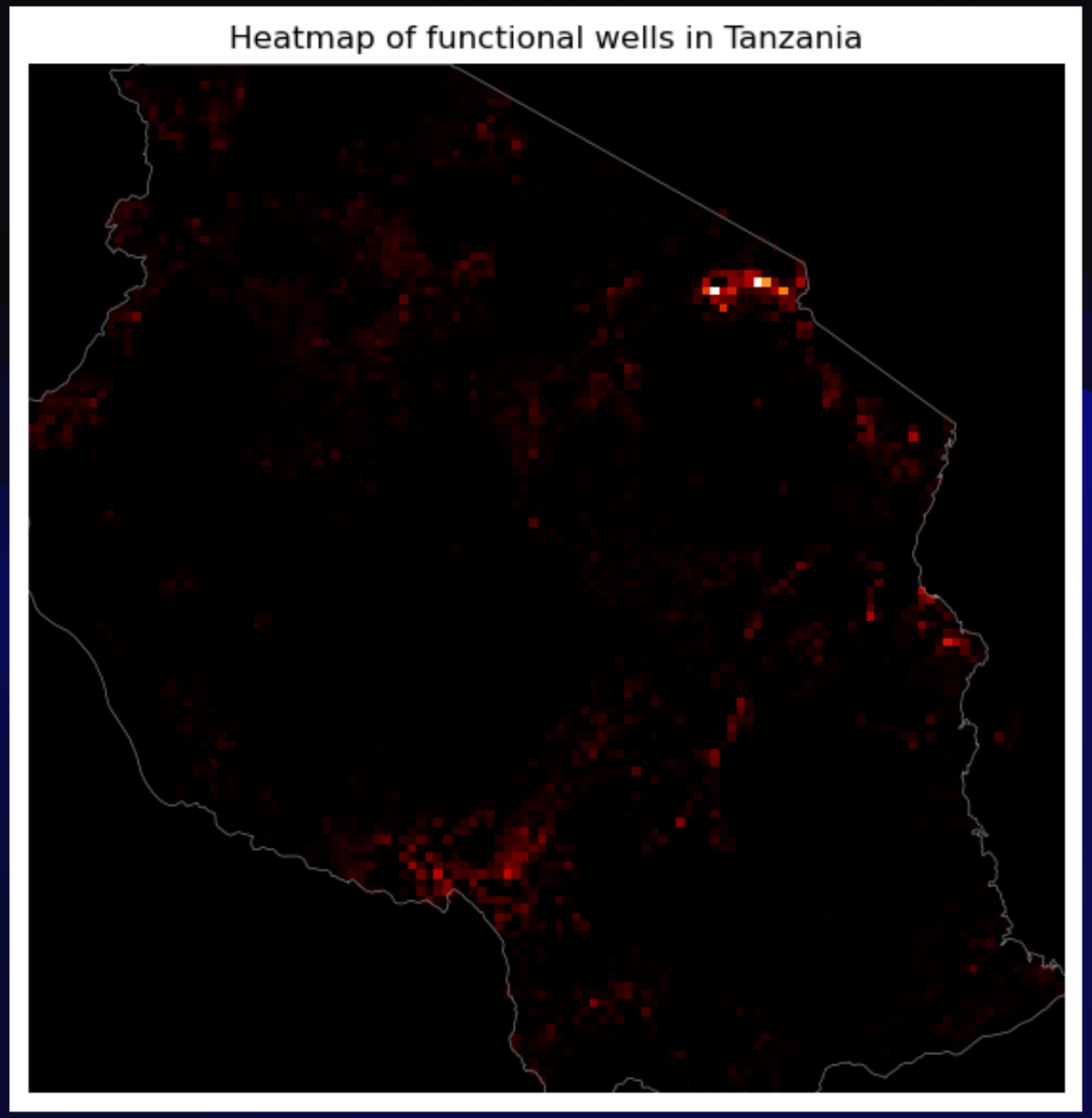


Recommendation #1

- Bear in mind that older wells are more likely to be dysfunctional

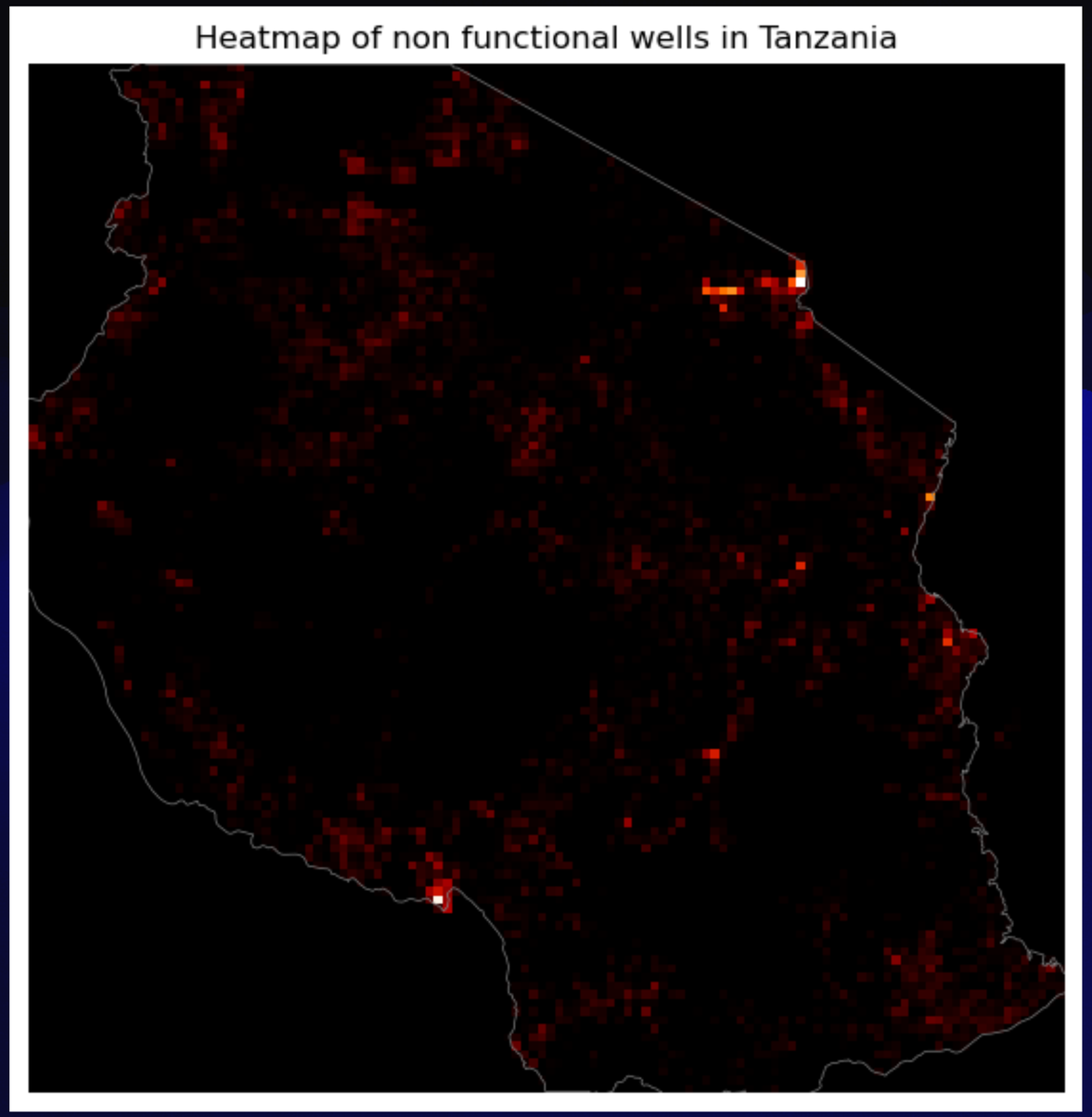
Heat map for functional wells in Tanzania

- Several hotspots along the coast



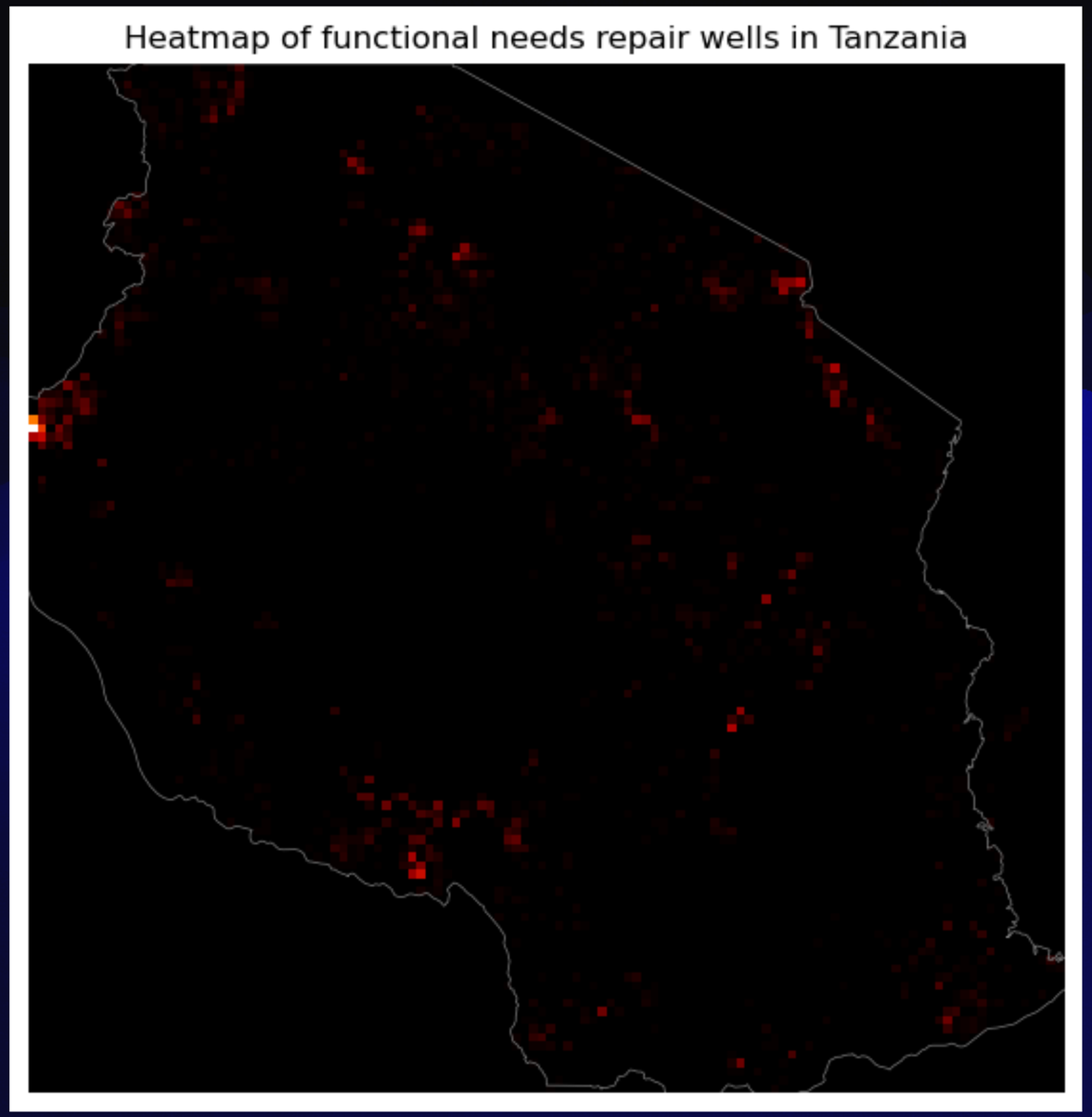
Heat map for non functional wells in Tanzania

- More evenly distributed across the country
- Major hotspots remain



Heat map for functional needs repair wells in Tanzania

- Less prevalent
- Still several hotspots

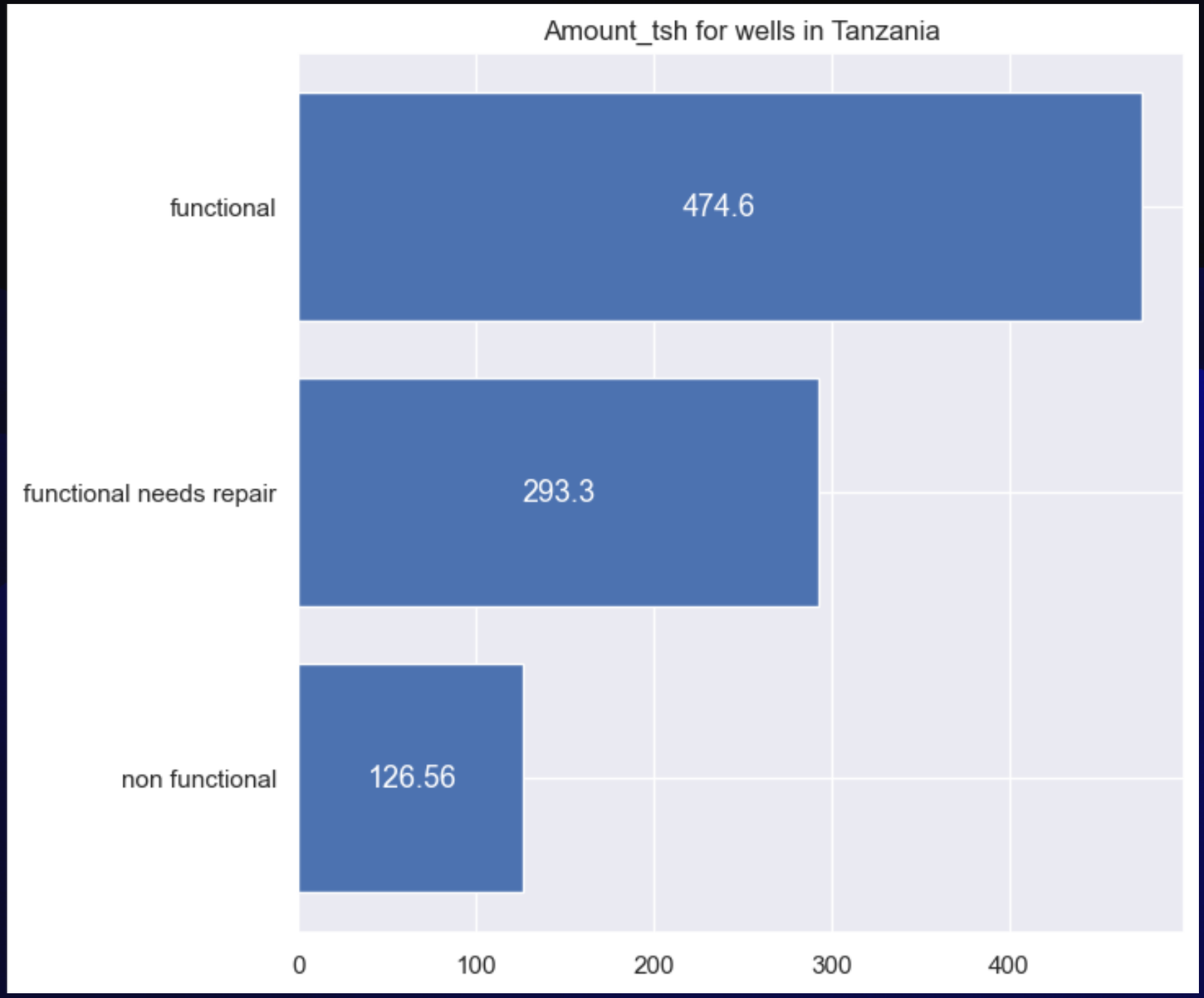


Recommendation #2

- Consult a heat map to better allocate resources

Average amount of water available to each status category

- More dysfunctional wells have less water available to them



Recommendation #3

- Prioritize wells that have less water available to them

Recommendation #4

- Prioritize non-functional over functional-needs-repair wells
- Numerous models were unable to accurately predict functional-needs-repair wells, suggesting an ill-defined category

Thank You!

- All questions are welcome.