PHASE 3 PRESENTATION

Subject: Tanzanian Water Wells

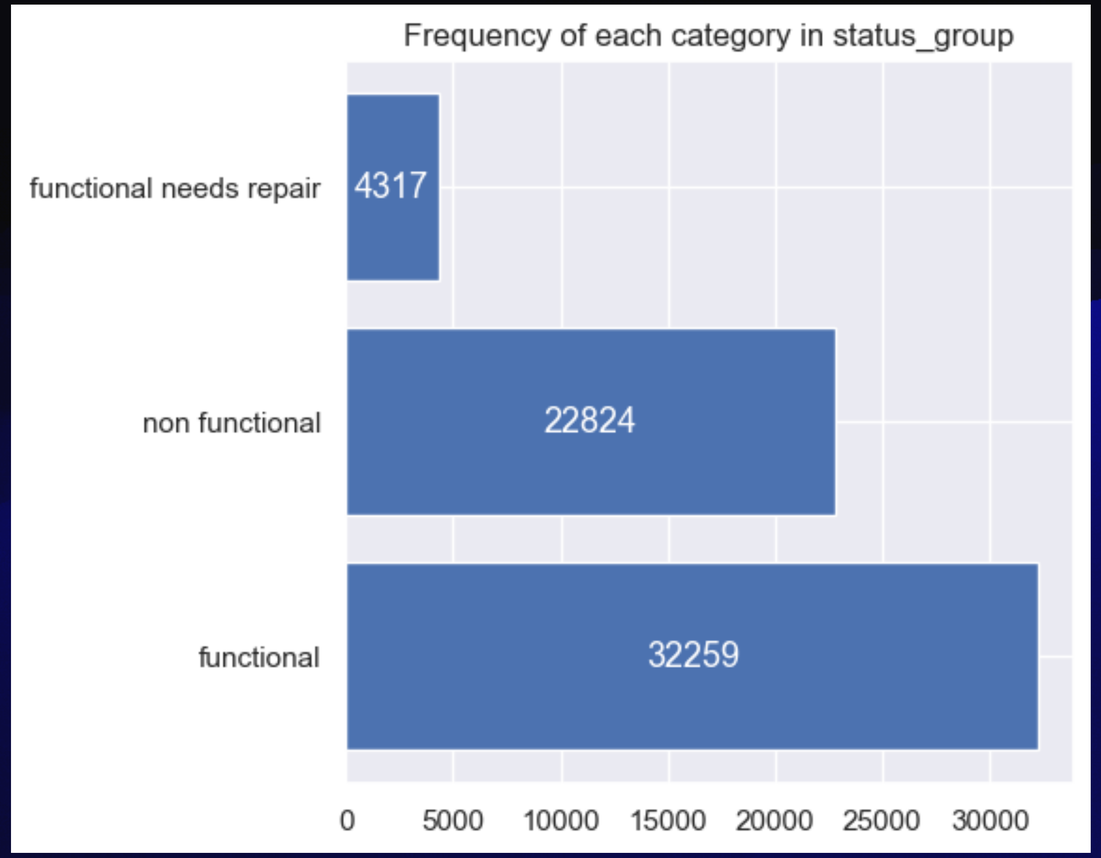
Angelo M. Turri

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

- 40 features is too many, we cannot keep them all
- Several variables aren’t suitable for our models
  - Do not correlate with target variable (e.g., id column)
  - Correlate with other features (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

Numeric Features

- amount_tsh: Amount of water available to each waterpoint
- vicinity_amount_tsh: Average amount_tsh for water points in the vicinity
- gps_height: Altitude of the well
- longitude: GPS coordinate
- latitude: GPS coordinate
- num_private: number of private waterpoints available to the owner
- population: Population around the well
- vicinity_population: Average population for water points in the vicinity
- construction_year: The year each waterpoint was constructed

Categorical Features

- lga (Geographic location): 124 categories.
- public_meeting (True/False): 3 categories.
- extraction_type (The kind of extraction the waterpoint uses): 18 categories.
- management (How the waterpoint is managed): 12 categories.
- payment (How people pay for water at the waterpoint): 7 categories.
- water_quality (The quality of the water): 8 categories.
- quantity (The quantity of water each waterpoint provides): 5 categories.
- source (The source of the water): 10 categories.
- waterpoint_type (The kind of waterpoint): 7 categories.

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

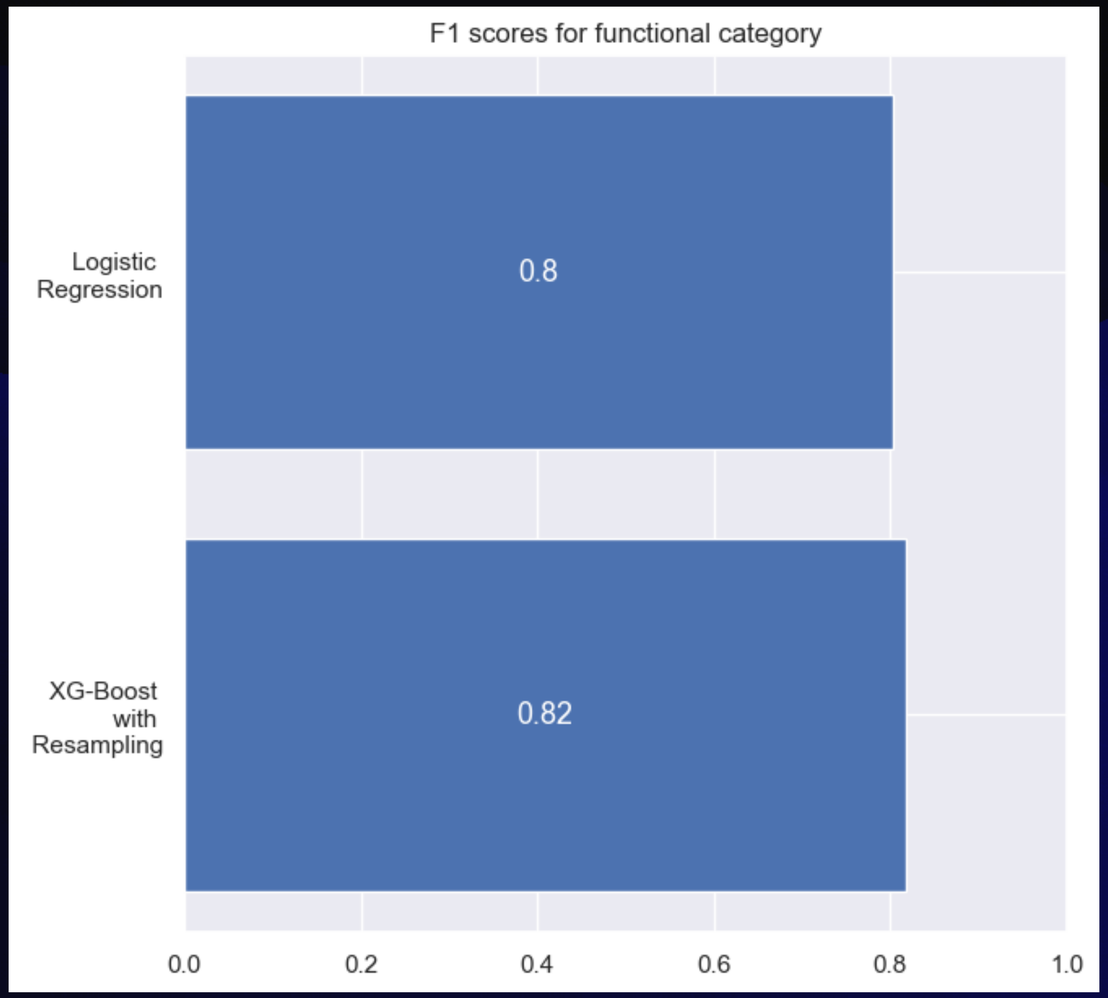
Estimators used

- 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: XG-Boost with 6,860 resampled instances of the “functional needs repair” category

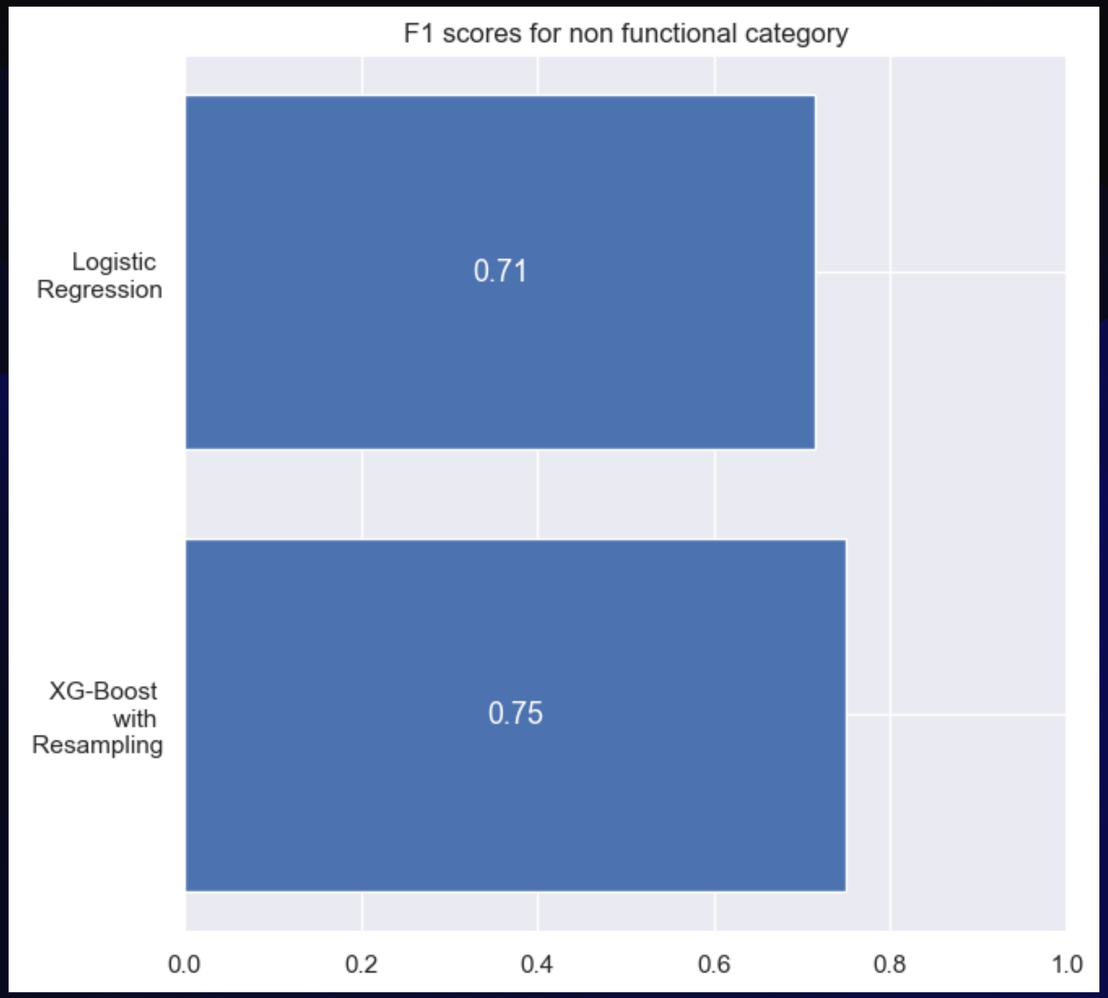
SMOTE explanation

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

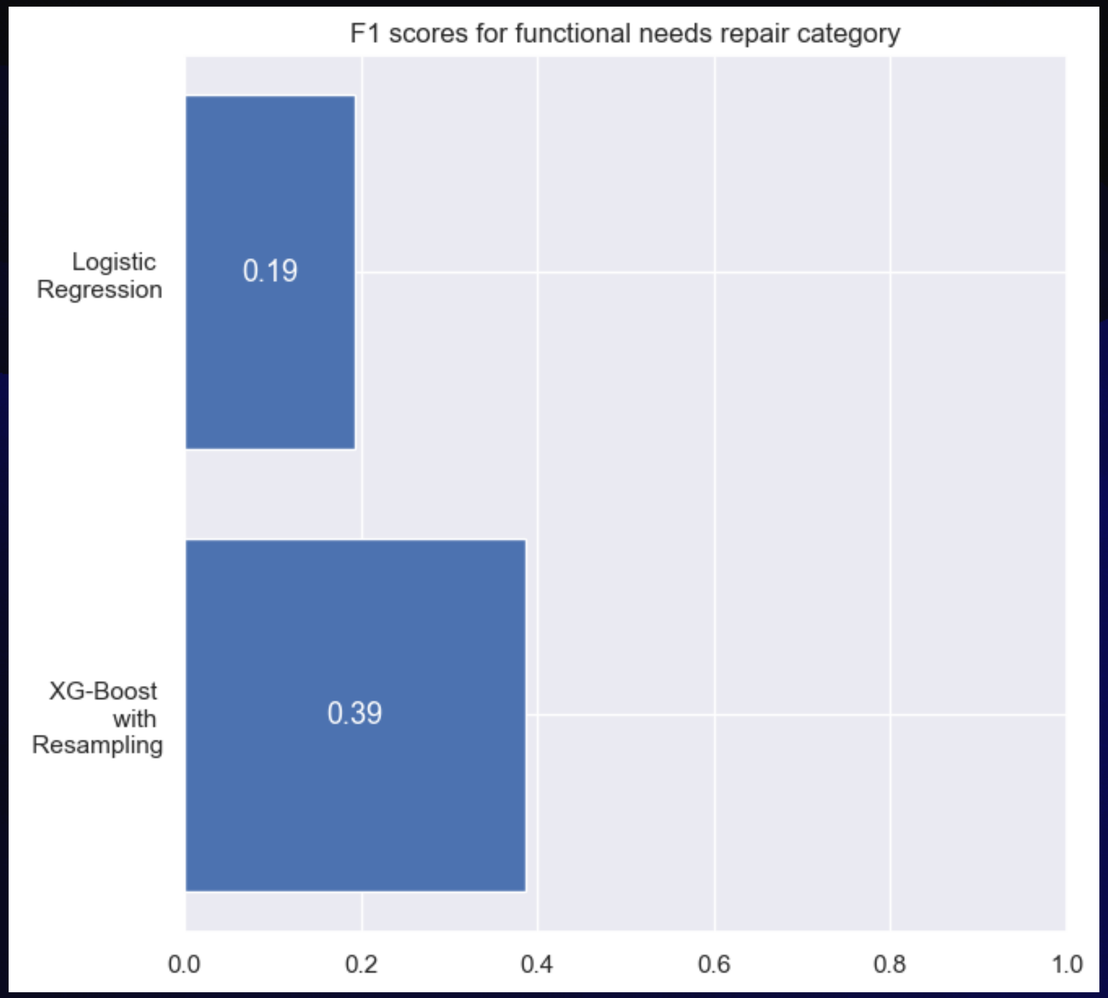
Comparing F1 scores



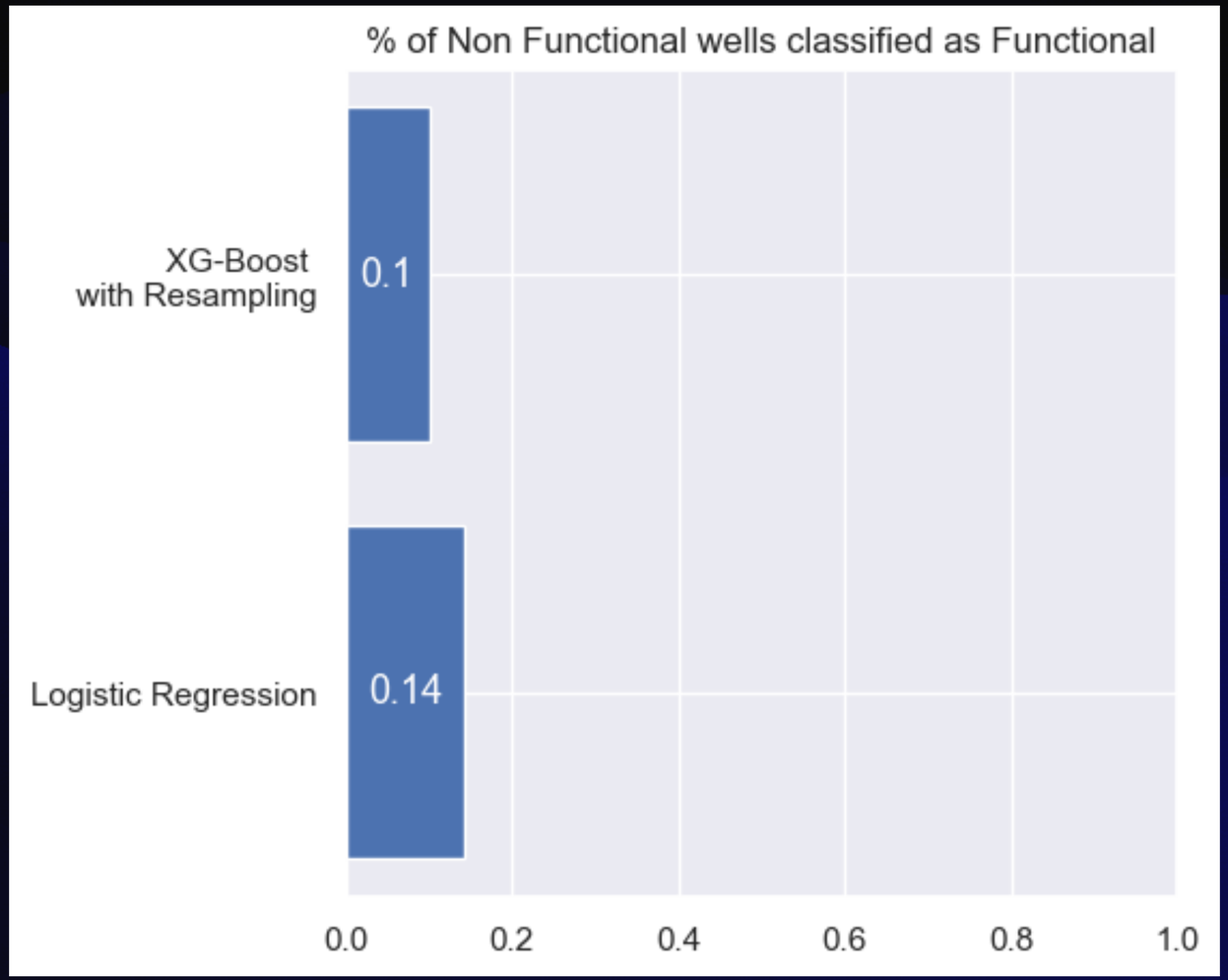
Comparing F1 scores



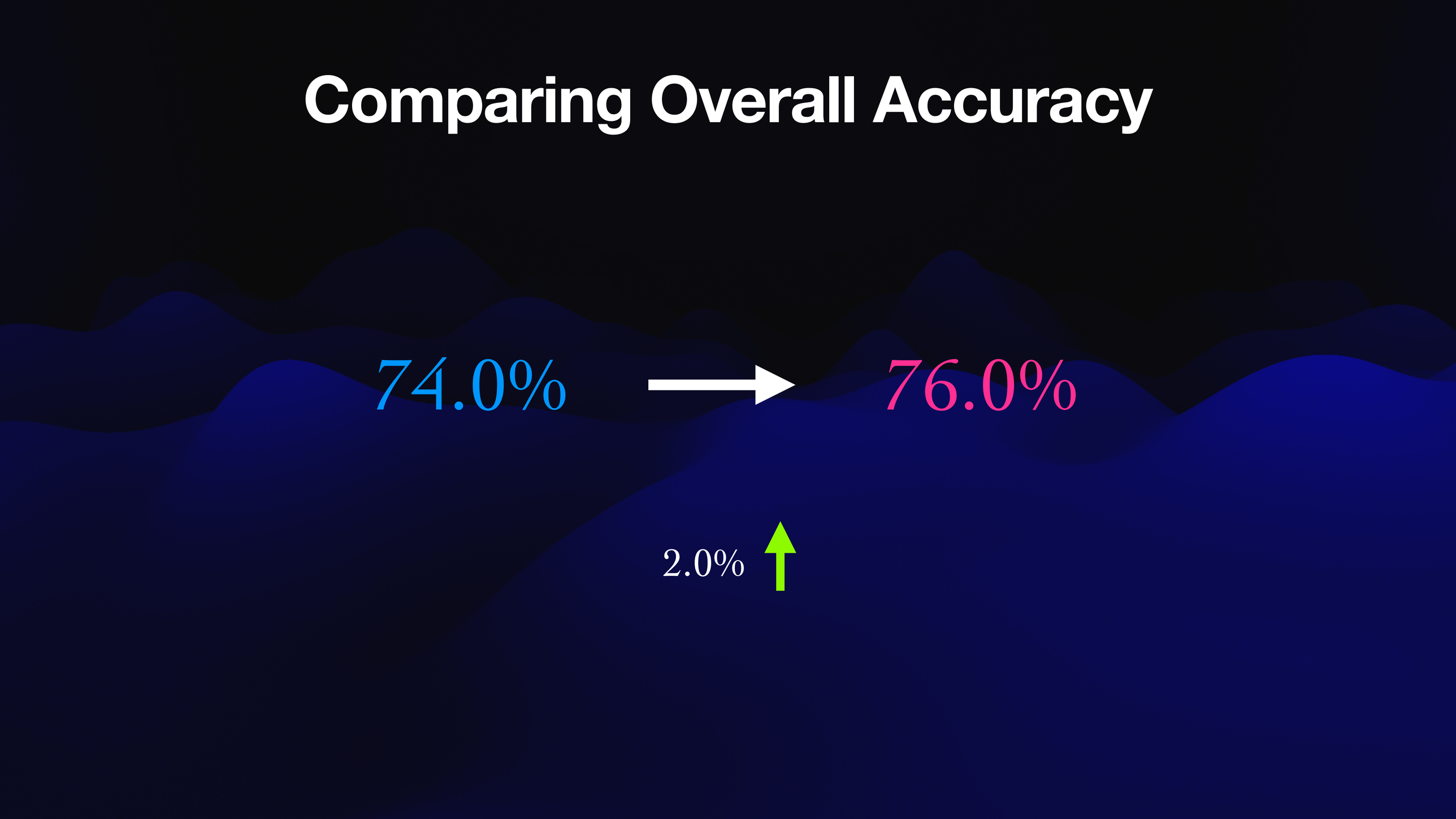
Comparing F1 scores



Comparing Error Scores



Comparing Overall Accuracy

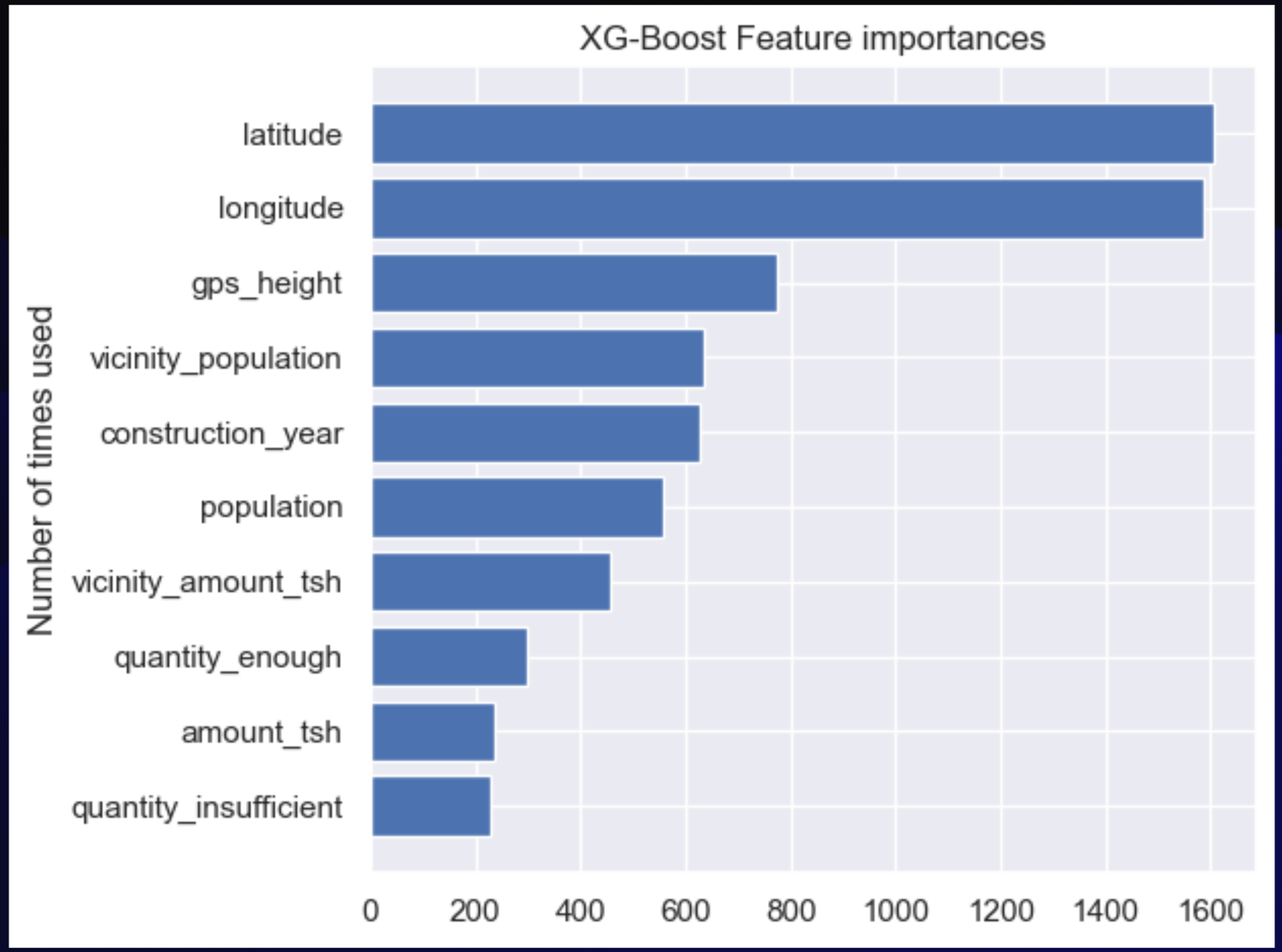


Providing the Requested Predictions

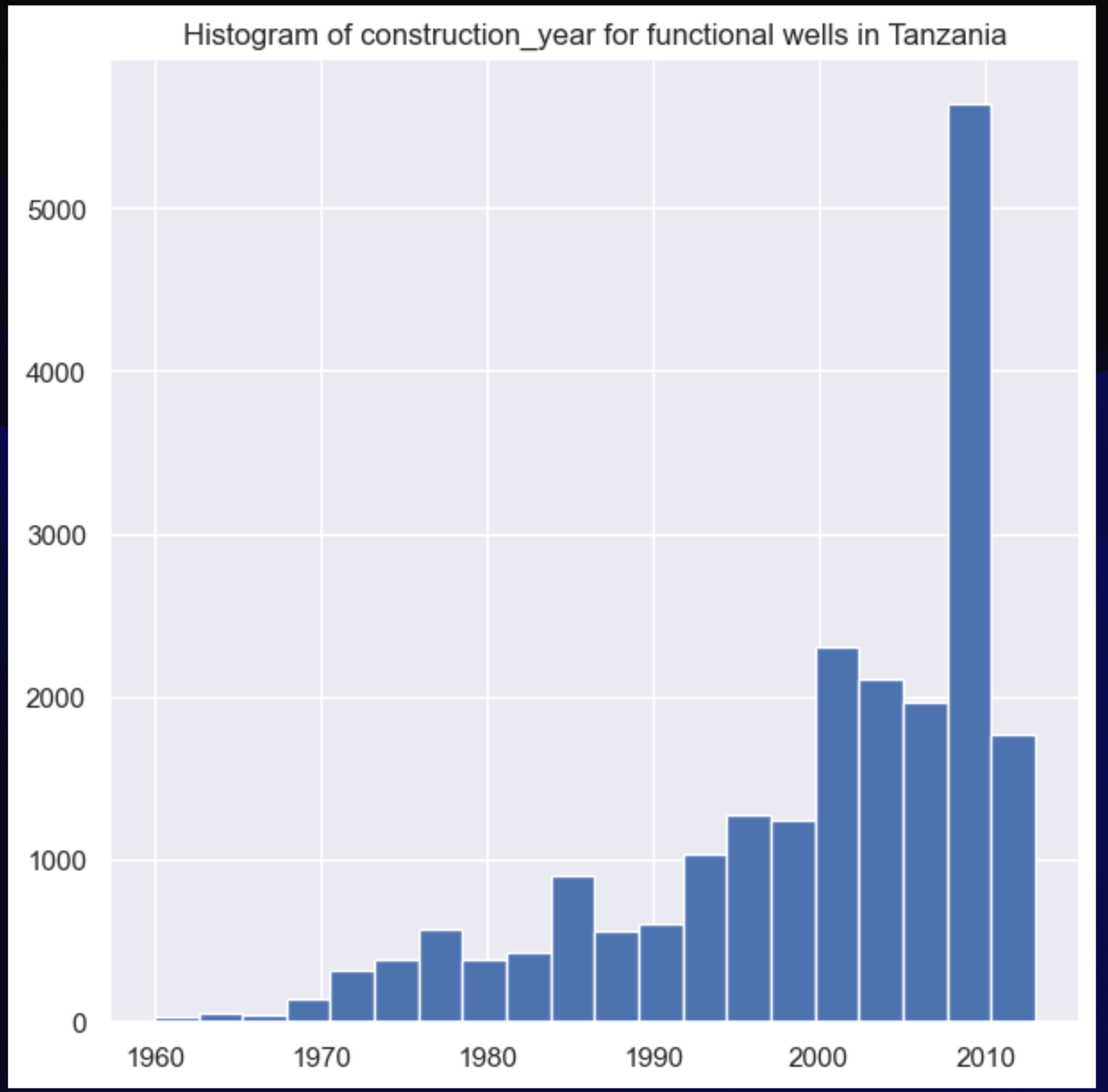
- This is where I would provide the requested predictions if necessary
- If it matters, I submitted them on drivendata.org, which is where the original competition is being hosted, and I placed roughly #4600 out of almost #16,000 participants.

Feature Importances

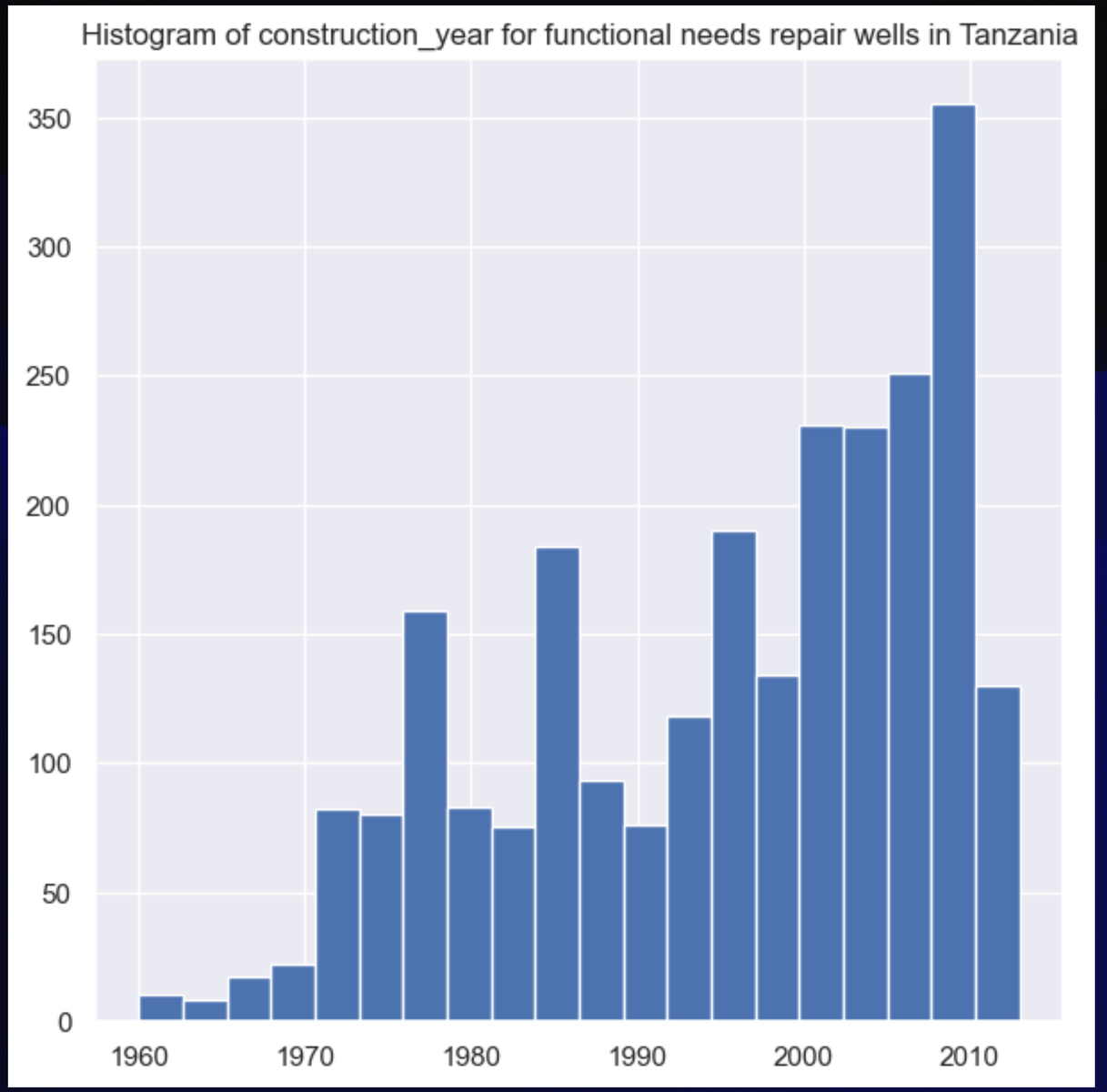
- The variables that our model deemed most important



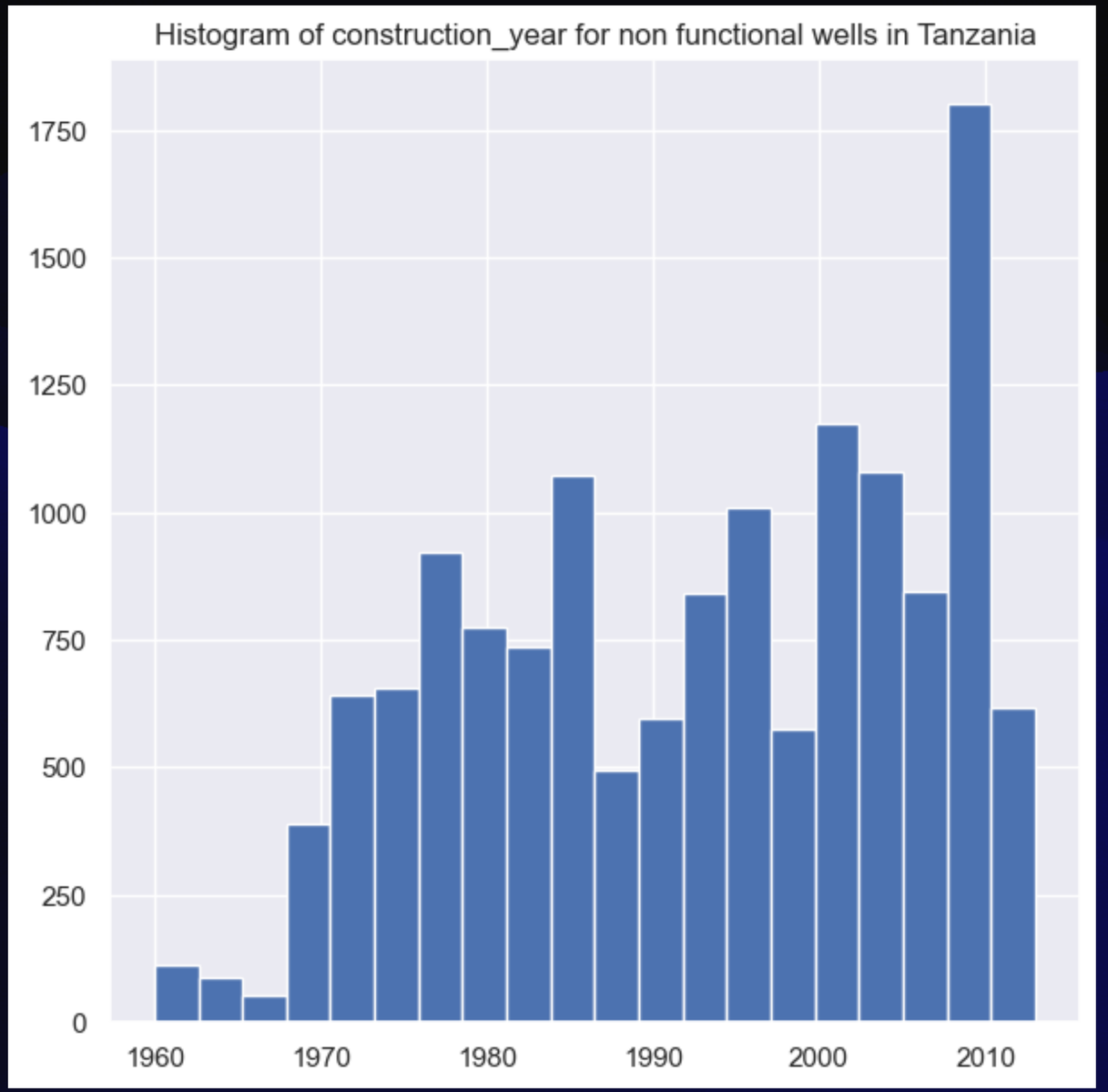
Construction year distribution



Construction year distribution

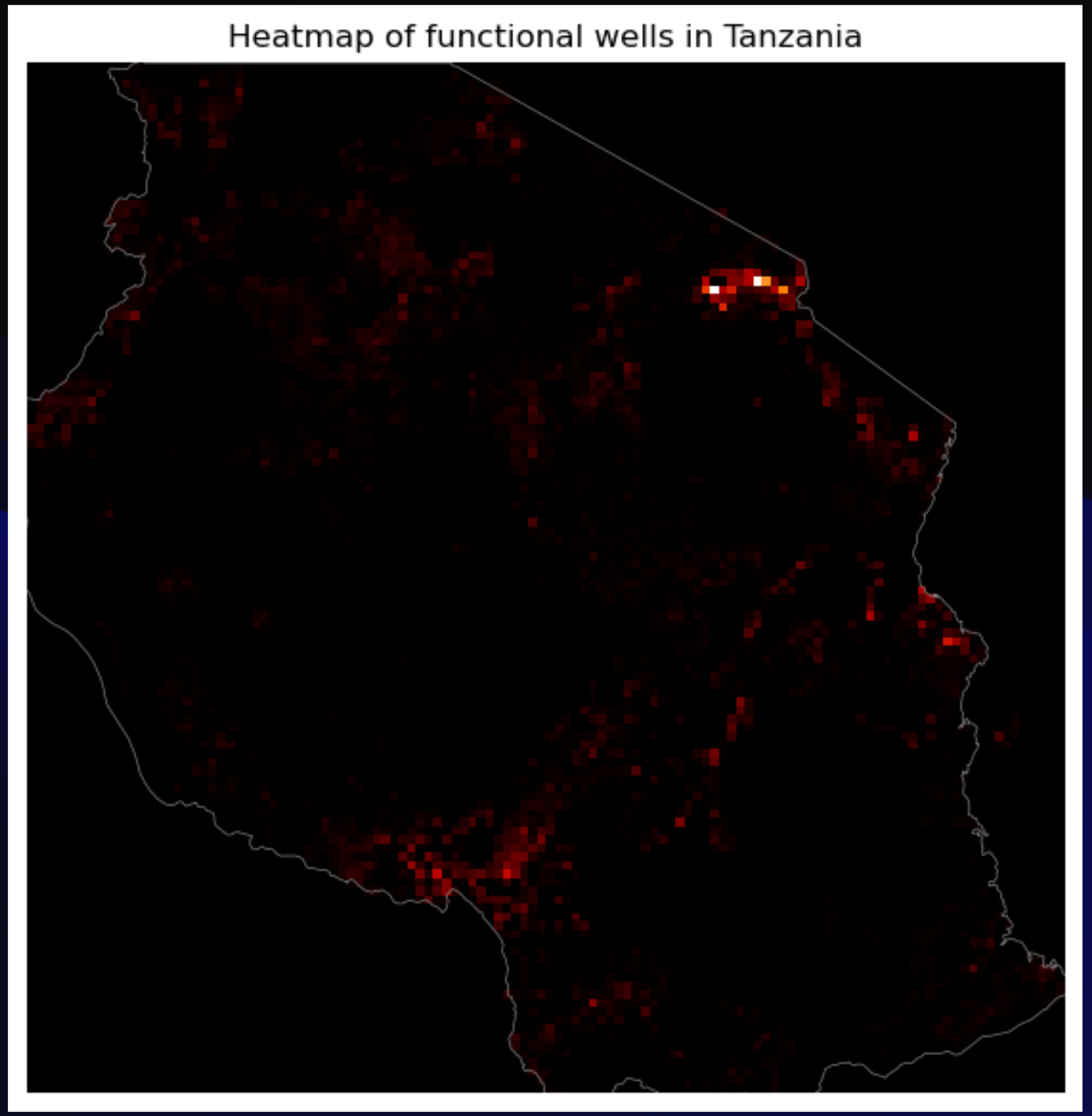


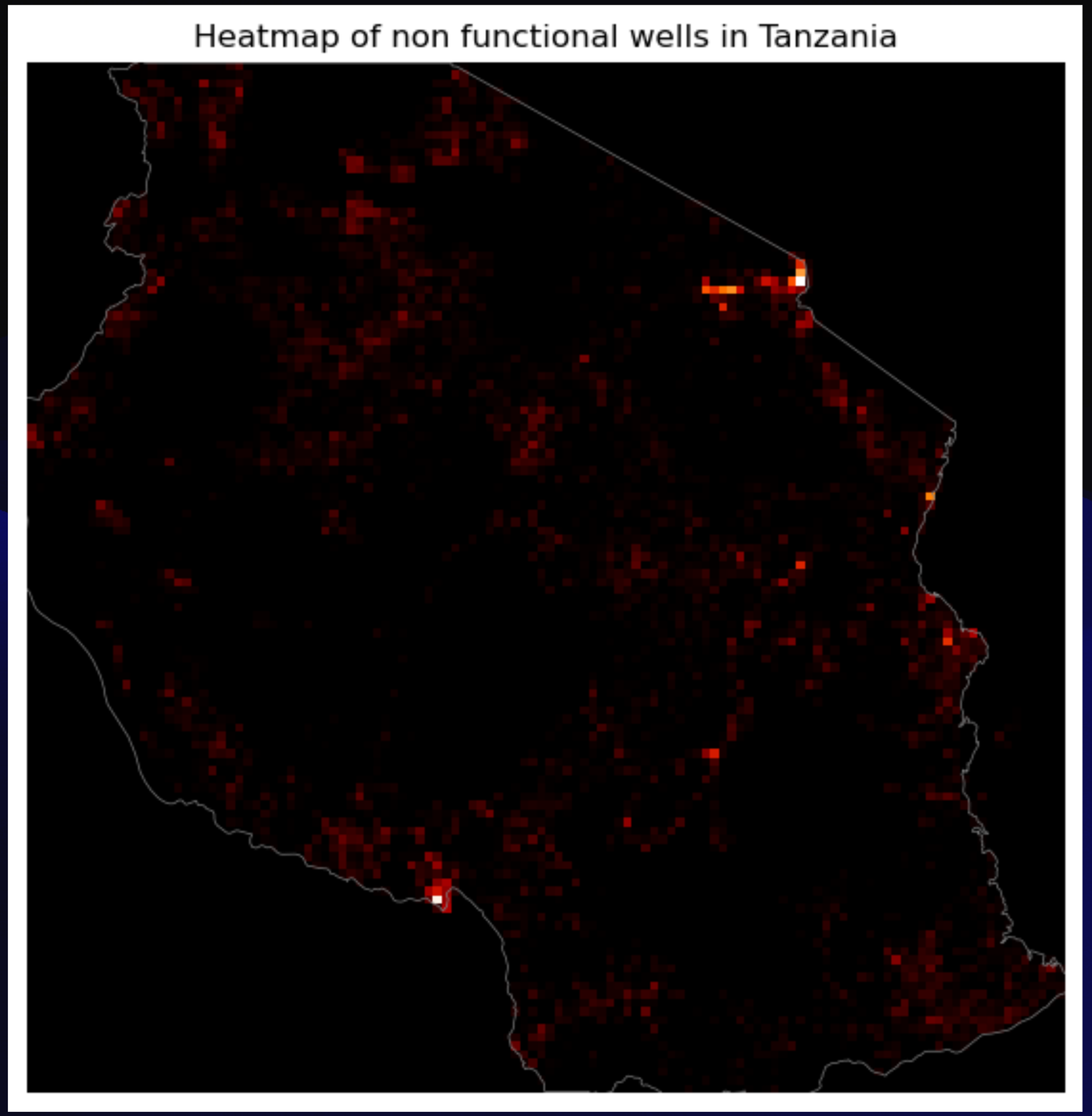
Construction year distribution



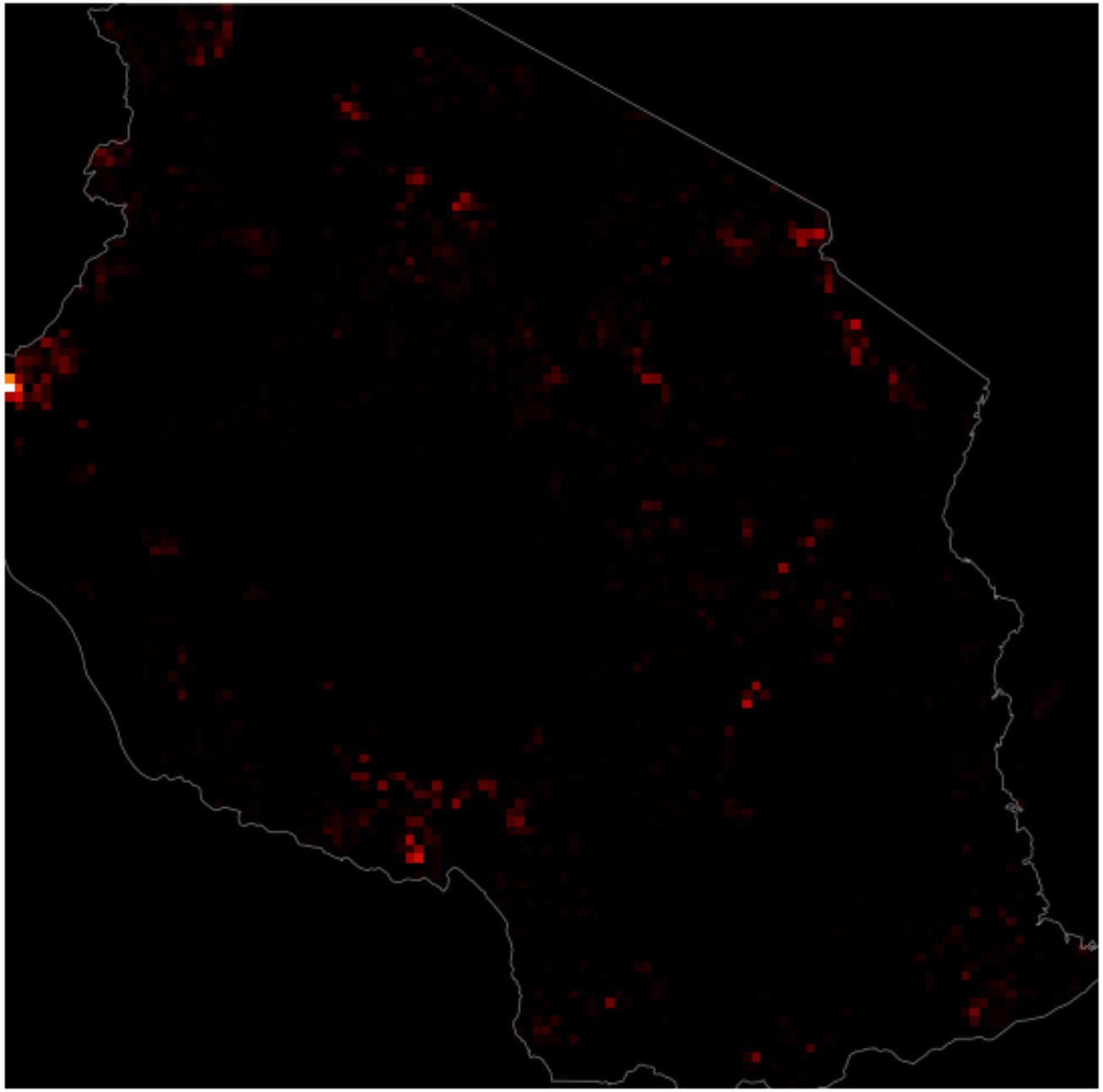
Recommendation #1

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


Heatmap of functional wells in Tanzania

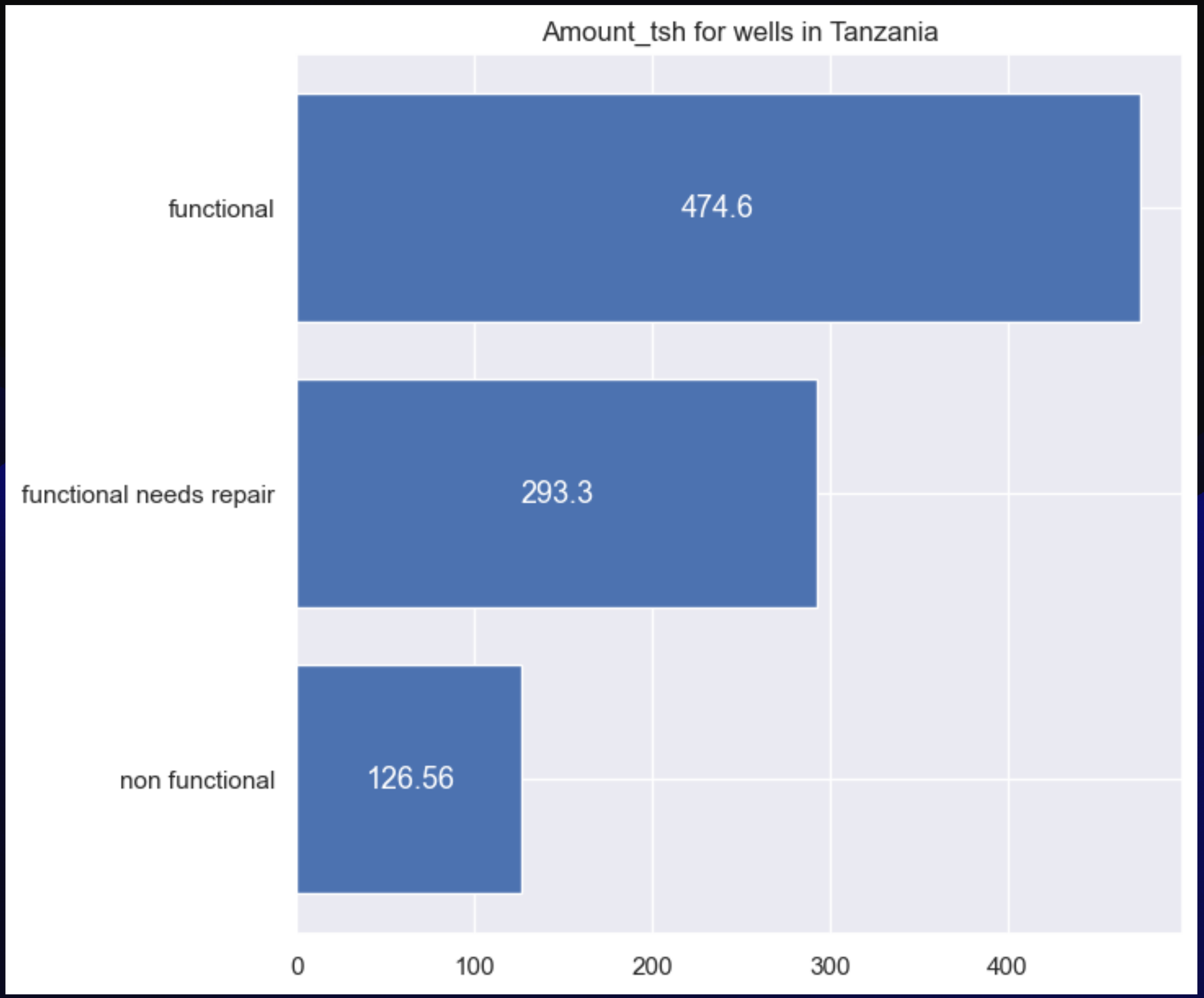

Heatmap of non functional wells in Tanzania

Heatmap of functional needs repair wells in Tanzania



Recommendation #2

- We suggest that the charity organization consult a heatmap of where non functional and functional-needs-repair wells are concentrated to better allocate their resources.


Amount_tsh for wells in Tanzania
functional
474.6
functional needs repair
293.3
non functional
126.56
0
100
200
300
400

Recommendation #3

- We suggest that the charity organization prioritize wells that have less water available to them, since these wells are more likely to be non functional or in need of repair.

Recommendation #4

- Finally, we suggest that the charity organization prioritize non functional over functional-needs-repair wells. Despite making improvements in predicting functional-needs-repair wells, we were unable to achieve satisfactory accuracy in this category. Our best model only idenfitied 43% of all wells in this category, and when the model predicted such a well, it was only correct only 38% of the time.
- This suggests that the category is ill-defined, and a well in "need of repair" could be almost totally fine, or almost completely broken and just barely functional.

Thank You!

- All questions are welcome.