UN-Water: Resource Efficiency

Evan Rosenbaum and Matthew Silver
DataScience
DTS-LIVE-051324-P3

Summary

We assessed data from the Tanzanian Ministry of Water to determine what indicators lead to a well being in need of attention.

- What impact will it have if we can predict well that need attention?
- What are some of the leading features that lead to well failure or being in need of repair?
- How can UN water enhance water availability and maintain efficiency across Tanzania?

Outline

- Business Problem
- Data and Methods
- Results
- Conclusions

Business problems

- Our client - The UN-water wants us to determine what factors lead to a well being in need of attention.
    - **Goal**: Find the leading indicators that requires a well to need attention.

Questions:

1. What preprocessing steps do we need to take to create an effective predictive model?
2. Which type of predictive model gets the best results?
3. What hypertuning techniques are used to tune the model?
4. How does our model add value?

Data Understanding

Dataset:

- Data from the Tanzanian Ministry of Water regarding the operating condition of a waterpoint (well) for each record in the dataset.
  - 59,400 rows and 39 columns

Data Understanding

Scoring Metric:

- Recall over accuracy
  - Improving recall works to reduce predictions where wells that need attention are incorrectly predicted as not needing attention.
- Chosen as it enhances the model's effectiveness in prioritizing well maintenance.

Data Understanding

Pre-processing:

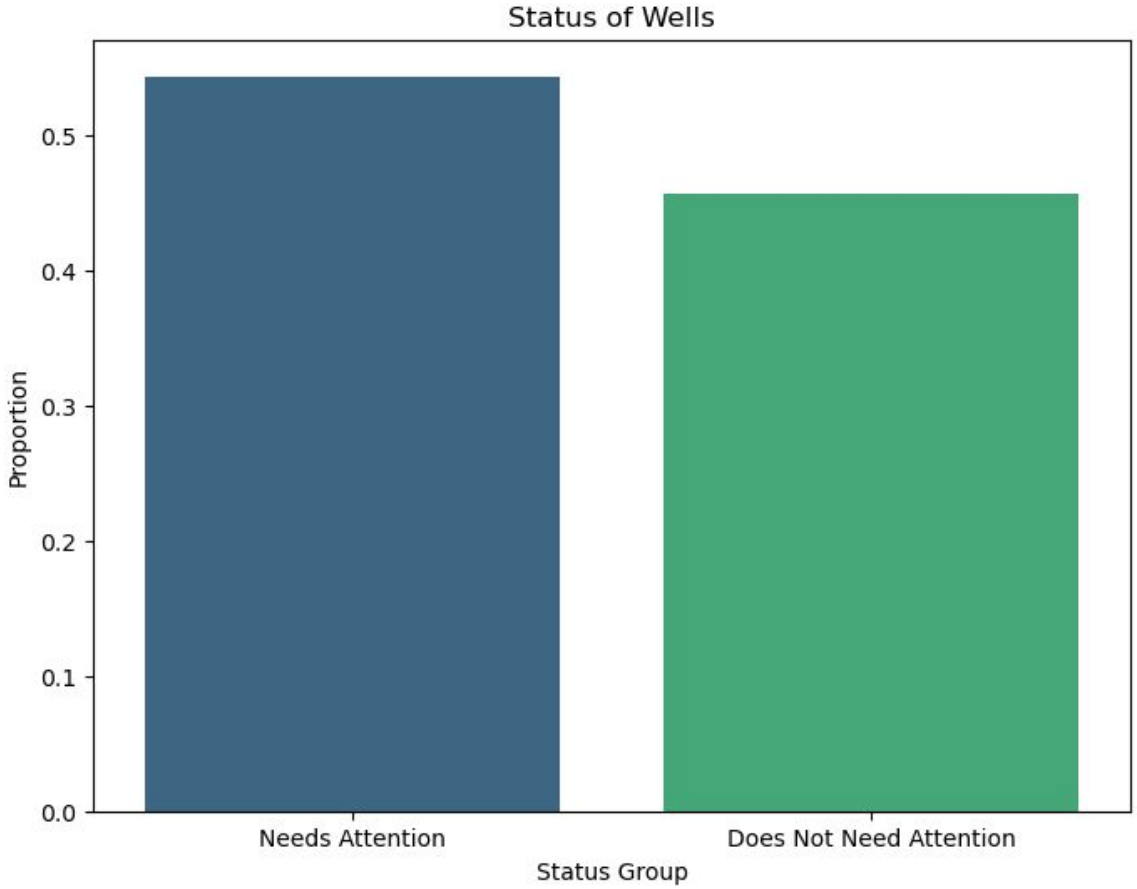
- Cleaning:
  - Removed illogical values.
    - Wells that are 117,000 meters deep.
    - Longitude and Latitude at 0°N, 0°E.
  - Handled typos and formatting issues.
    - “cebtral government”
    - “unisef”
  - Fill categorical NaNs with ‘unknown’.
  - No missingness in numerical features.
- Dropping features:
  - Too similar features.
  - Not documented features.
  - L1 Penalty/VIF determined unimportant features.

Data Understanding

Target Variable:

- Classes:
  - There were three classes inside of the target variable.
    - Functional
    - Non-Functional
    - Functional needs repair
  - We turned the classification problem into a binary classification.
    - Does not need attention (for functional wells)
    - Needs attention (for non-functional and functionally needs repair wells)
- Prediction Goal:
  - Inform the ministry in advance as to which pumps are most likely to be in need of attention so that they can prioritize and triage the well.

Target Variable Distribution



Model Results

- We ran two models with varying feature selection and sampling techniques.
- Dimensionality reduction resulted in poorer scores.
- VIF decision trees were our best models so we hypertuned them.

	Model	Feature Selection	Sampling Technique	Recall-Score
15	DecisionTreeClassifier	VIF	oversample	0.770180
17	DecisionTreeClassifier	VIF	smote	0.768383
16	DecisionTreeClassifier	VIF	undersample	0.767511
6	LogisticRegression	L1_Penalty	oversample	0.704815
0	LogisticRegression	base	oversample	0.704379
7	LogisticRegression	L1_Penalty	undersample	0.703835
1	LogisticRegression	base	undersample	0.703290
18	LogisticRegression	dimensionality_reduction	oversample	0.703072
19	LogisticRegression	dimensionality_reduction	undersample	0.701656
8	LogisticRegression	L1_Penalty	smote	0.701493
2	LogisticRegression	base	smote	0.701166
20	LogisticRegression	dimensionality_reduction	smote	0.699586
13	LogisticRegression	VIF	undersample	0.683299
12	LogisticRegression	VIF	oversample	0.680630
14	LogisticRegression	VIF	smote	0.671370
5	DecisionTreeClassifier	base	smote	0.539384
23	DecisionTreeClassifier	dimensionality_reduction	smote	0.539329
11	DecisionTreeClassifier	L1_Penalty	smote	0.539275
10	DecisionTreeClassifier	L1_Penalty	undersample	0.531103
22	DecisionTreeClassifier	dimensionality_reduction	undersample	0.531103
4	DecisionTreeClassifier	base	undersample	0.531049
9	DecisionTreeClassifier	L1_Penalty	oversample	0.515142
3	DecisionTreeClassifier	base	oversample	0.515033
21	DecisionTreeClassifier	dimensionality_reduction	oversample	0.515033

Hypertuned Model Results

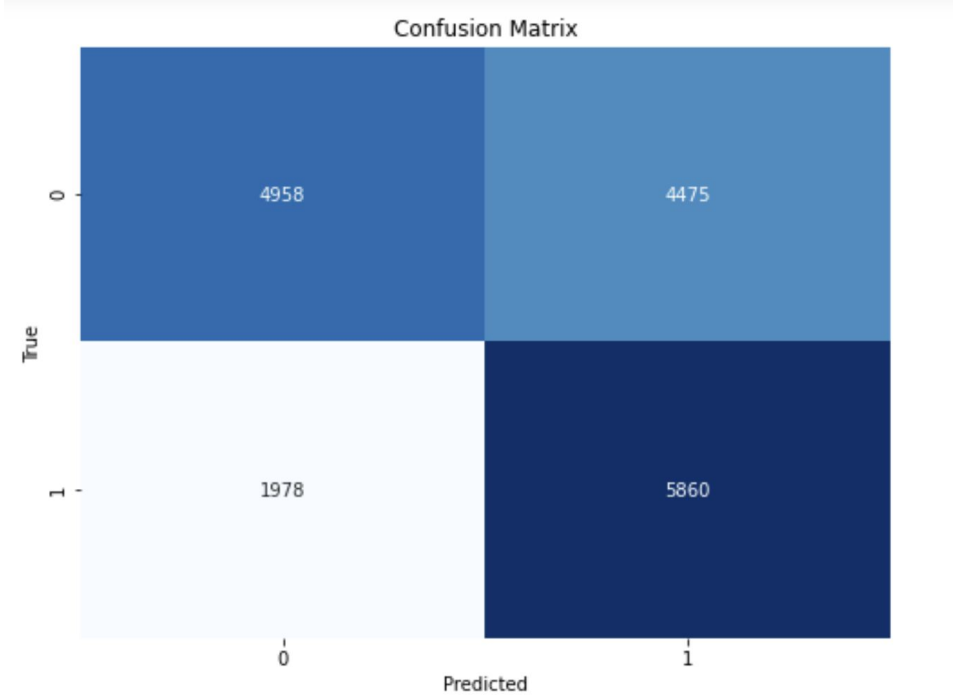
	Model	Feature Selection	Sampling Technique	Original Recall-Score	Hypertuned Best Recall Score
15	DecisionTreeClassifier	VIF	undersample	0.768383	0.749024
17	DecisionTreeClassifier	VIF	oversample	0.770180	0.741414
16	DecisionTreeClassifier	VIF	smote	0.767511	0.740672

- All models decreased slightly.
  - Indicative of an overfit model.
- Oversampling vs Undersampling
  - Chose Oversampling as our best model.

Hypertuned Model Results

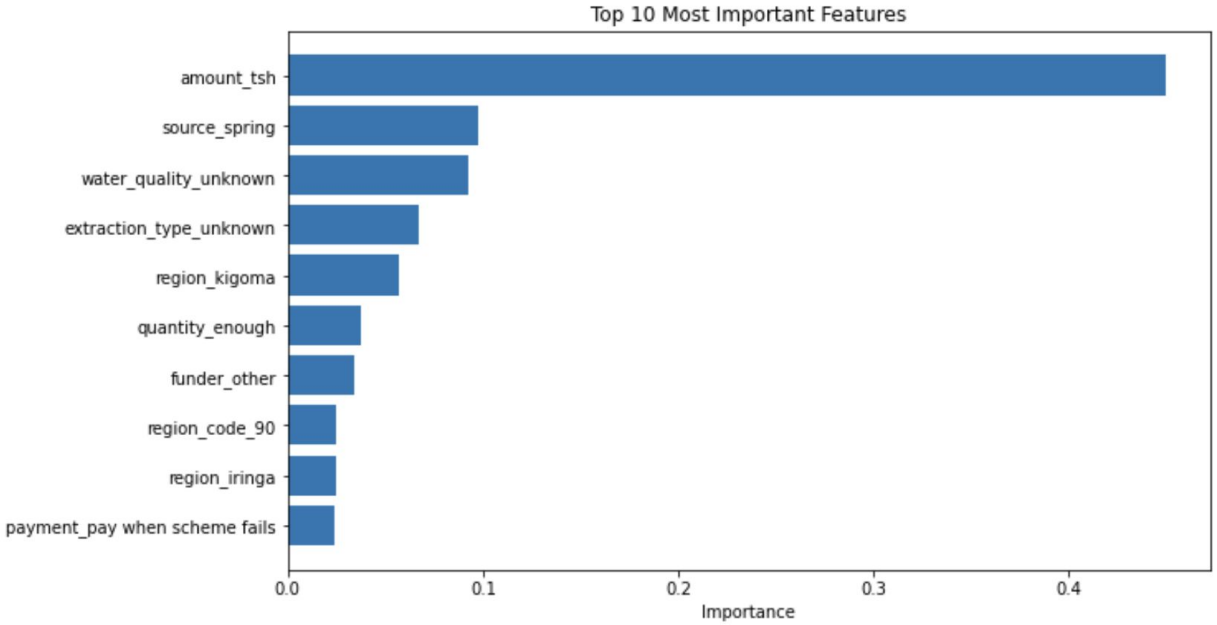
- 0 are wells that do not need attention.
- 1 are wells that do need attention.

- On the final test set, the oversampling model with VIF features had a 0.7476 recall score.



Most Important Features: Best Hypertuned Model

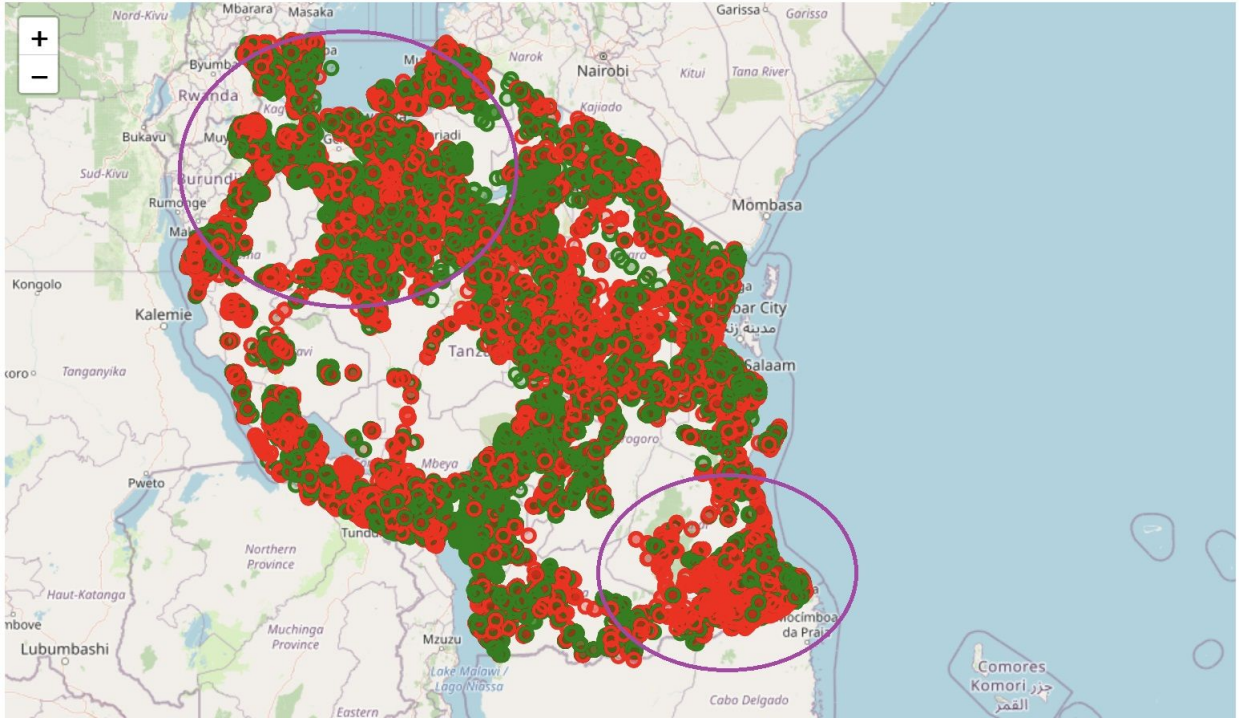
- Amount_tsh is the total static head or the amount water available to waterpoint.



Recommendations:

1. Prioritize wells in the northwest and southeast of the country.

- Red circles are wells that are in need of attention



Recommendations:

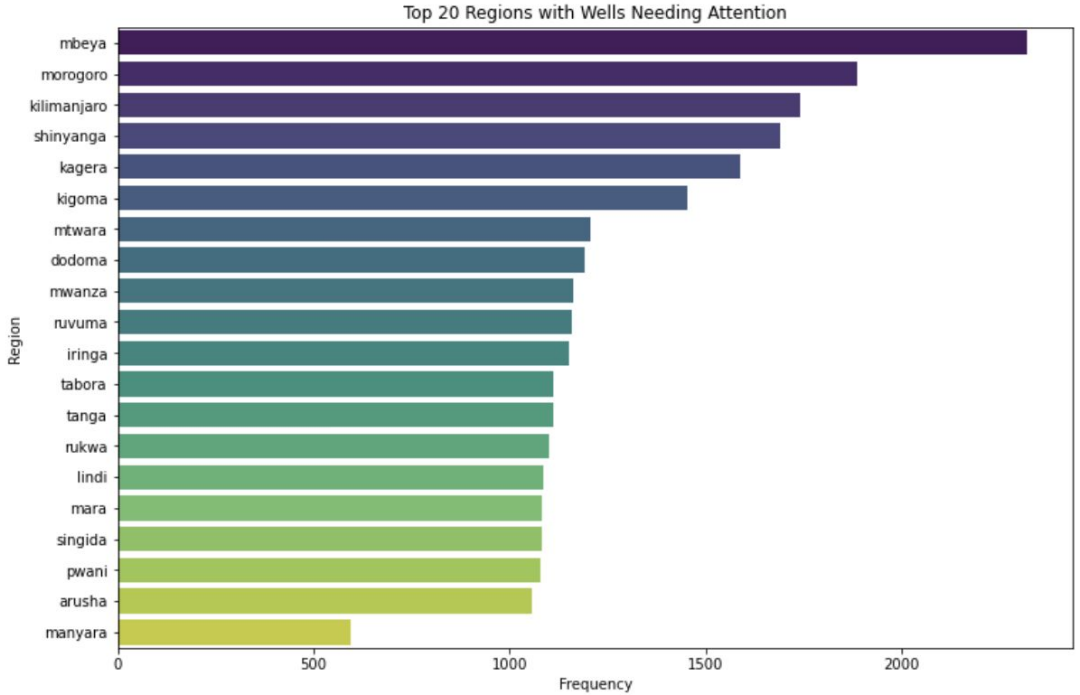
1. Prioritize wells in the northwest and southeast of the country.

Northwest:

- Kagera
- Kigoma
- Shinyanga
- Mwanza
- Mara

Southeast:

- Lindi
- Mtwara
- Iringa

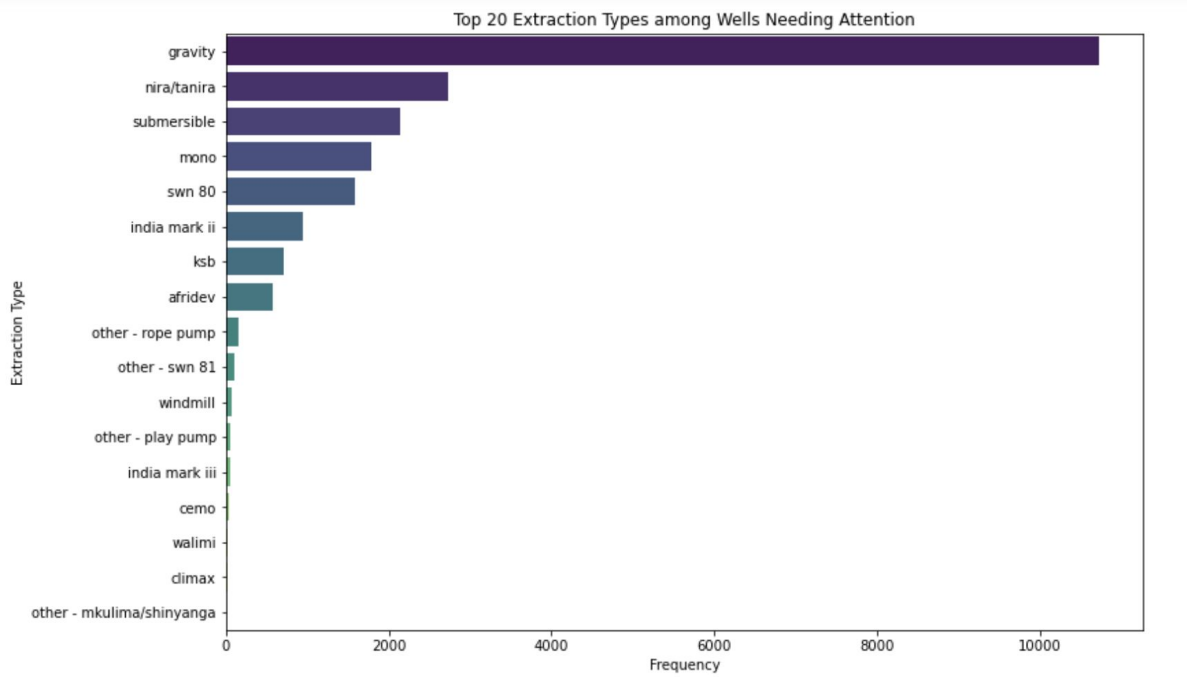


Recommendations:

2. For new wells, look to improve or use the latest version of gravity extraction pumps

- While counterintuitive, it appears as though the construction quality and materials used could be playing a role here.

- Gravity pumps should have fewer issues.



Predictive Recommendation

- All data provided was surveyed by GeoData Consultants Ltd. Our prediction model can be used in place of consultants, cutting costs.

- Benefit of the model
  - Can assist in efficient resource allocation and proactive maintenance planning.
    - Note
      - The model is based on historic results, will not be apt for use during natural disasters but can be used before-hand to identify high priority regions.

Next steps

- Additional data cleaning
  - Better grouping of the data
  - Test different encoding techniques on categorical data
- More feature selection techniques
- Additional classification models
  - Naive Bayes
- More robust hyperparameter tuning

Thank you!

Questions!

Contacts

Evan Rosenbaum


evanrosenbaum24@gmail.com


https://github.com/evan-rosenbaum1


https://www.linkedin.com/in/evan-rosenbaum/

Matthew Silver


silverma100@gmail.com,


https://github.com/silver032


https://www.linkedin.com/in/msilver/