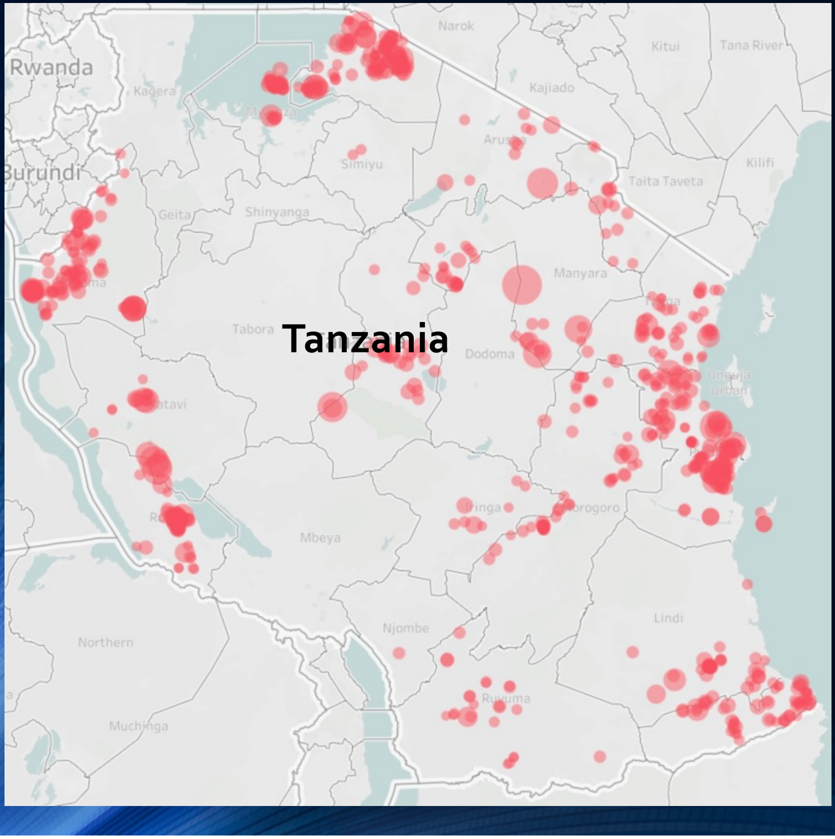
Predicting Water Well Failures Using Machine Learning

ANALYSIS BY BRIAN BENTSON

Predicting Well Failure Can Save Lives

Non-Functional wells that support over 1,000 citizens

Sized by population



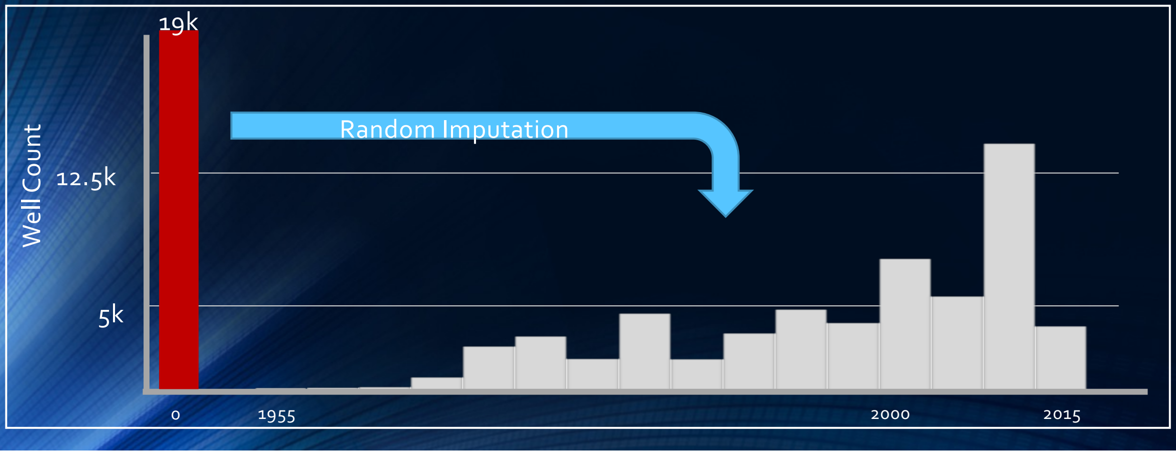
- Tanzania has almost 60,000 water wells in dataset with 45% not functional, leaving almost 7M people without a reliable water source
- Humans can only live up to 3 days without water
- Ability to predict water well failures and respond quickly can be the difference between life and death

Data Quality a Potential Issue

18K

- Almost 18,000 water wells with zeros for population, head, well elevation and construction year
- Recommendation: Improving data quality can drastically improve modeling performance

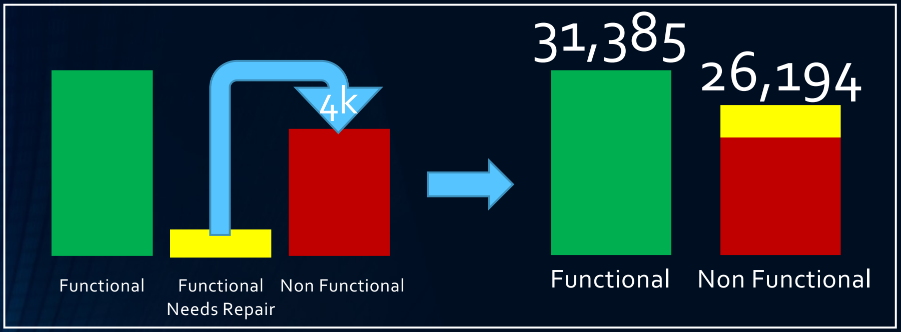
Distribution of Well Construction Year



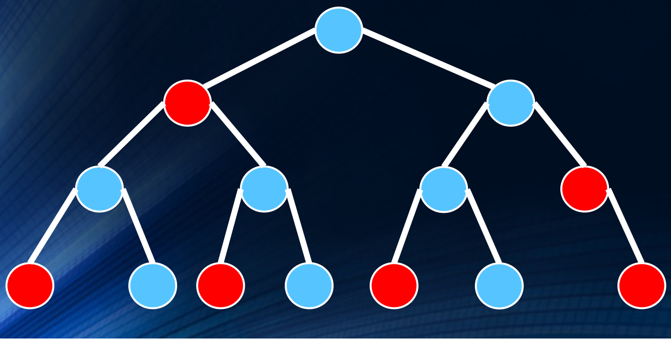
- Utilized random imputation to convert zeros to non-zero values
- Outliers and improbable values in the head that can skew results

Analysis Overview

Consolidation of Well Status



Random Forrest

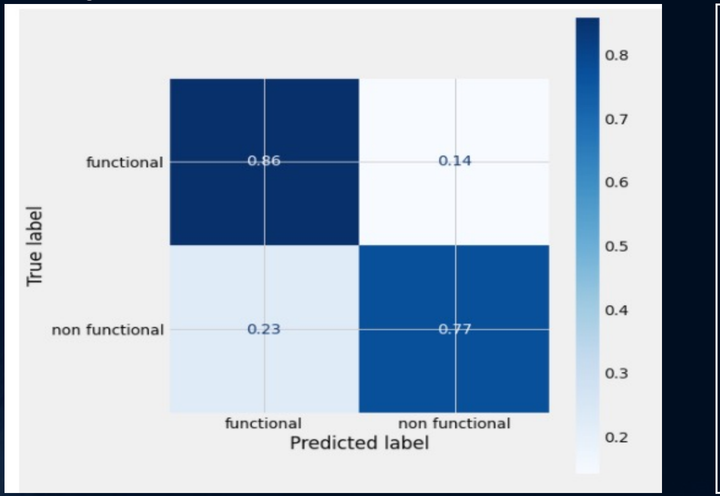


- Equated wells functioning at a reduced capacity as not functioning
- Prioritized finding all well failures, therefore accepting some false positives
- Focused on models with high interpretability to understand what drives well failures

Best Classification Modeling of Well Status

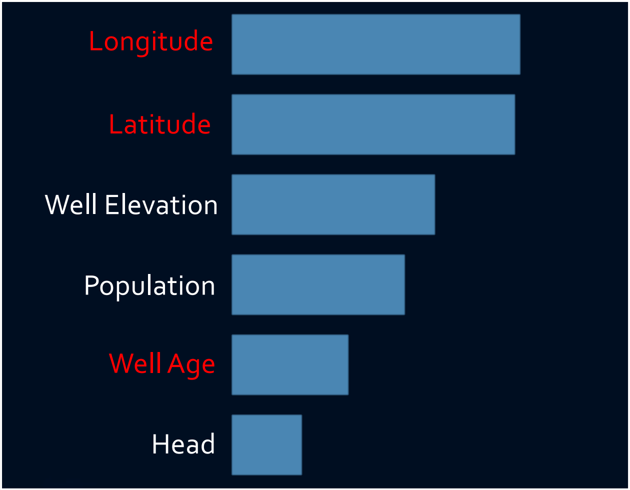
Random Forest Performance

Describing the final model



Random Forest Feature Importances

Sorted by Most Important

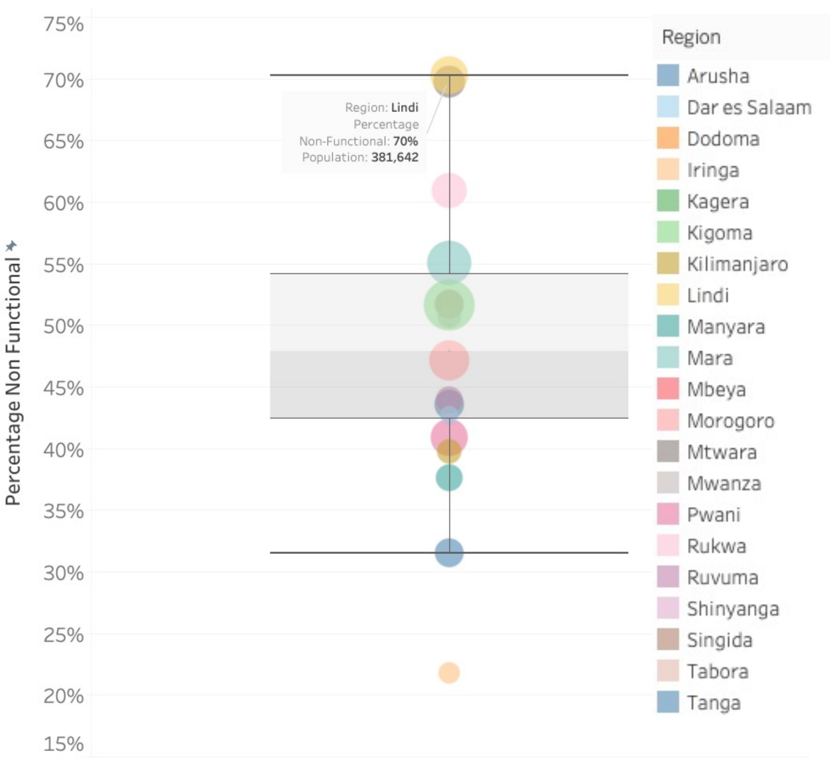


- Best model was a Random Forrest that could detect failures 77% of the time with an overall accuracy of 82%
- Location, well age and head are important for the model

Location Affects Reliability

Distribution Percentage Non-Functional by Region

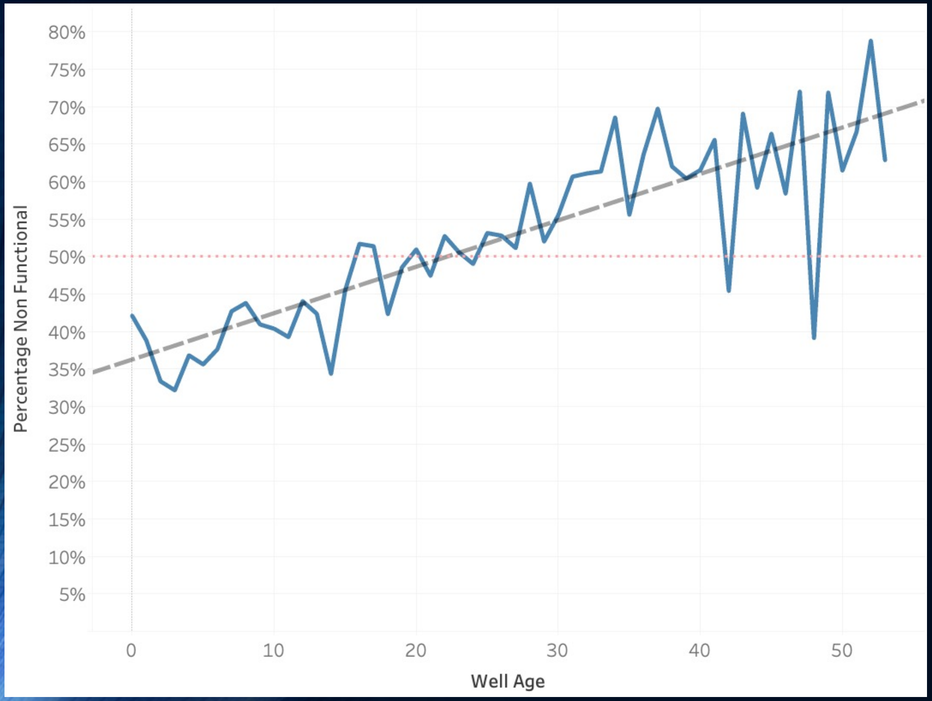
Sized by population



- High variability in well function across regions
- Many Regions with high failure percentage and high population
- Recommendation: Focus on regions which have historically high failure percentage

Well Age Negatively Affects Reliability

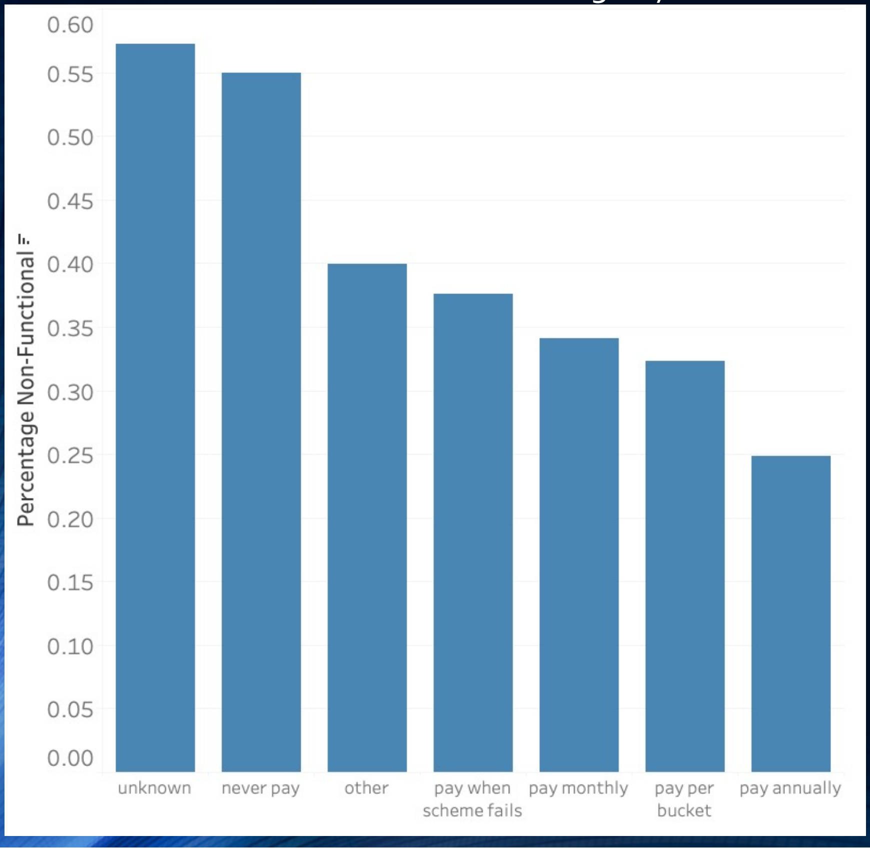
Percentage Non-Functional by Age



- As well ages increase, the percentage of non-functional wells increases
- If a well is over age 24-25, it is more likely that well will be non-functional rather than functional
- 2.5M people are supported by older wells (>24 years old)
- Recommendation: Focus maintenance on older wells to maintain supply of water

You Get What You Pay For

Distribution of Non-Functional Percentage by Water Cost



- While not specifically important for the random forest model, there is a clear trend between showing that if you pay for the water, the wells reliability is higher
- Recommendation: Focus on supporting the populations which cannot afford to pay for water

Recommendations & Next Steps

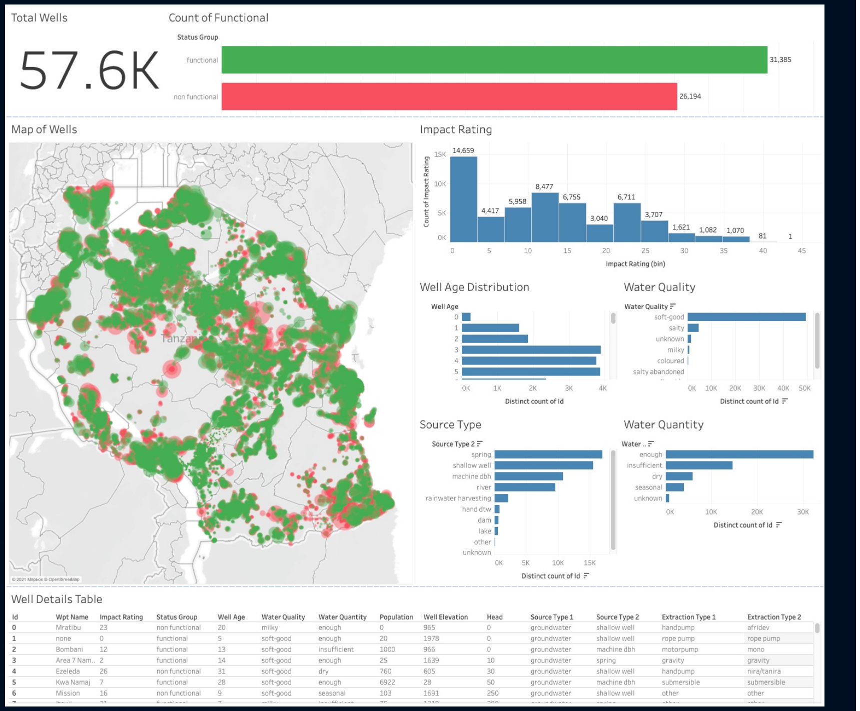
Recommendations

- Keep an eye on older wells as the likelihood of failure increases with age
- Develop support model for population areas that do not pay for water since analysis shows paying for water brings better reliability
- Utilize business insights tools to keep stakeholders up to date with key performance metrics
- Improve data governance to ensure better data quality and better predictions

Next Steps

- Use more sophisticated algorithms that may perform better at finding well failures but does not tell you why they are failing

Screenshot of BI Dashboard



Well Details Table

Id	Wpt Name	Impact Rating	Status Group	Well Age	Water Quality	Water Quantity	Population	Well Elevation	Head	Source Type 1	Source Type 2	Extraction Type 1	Extraction Type 2
0	Mratibu	23	non functional	20	milky	enough	0	965	0	groundwater	shallow well	handpump	afridev
1	none	0	functional	5	soft-good	enough	20	1978	0	groundwater	shallow well	rope pump	rope pump
2	Bombani	12	functional	13	soft-good	insufficient	1000	966	0	groundwater	machine dbh	motorpump	mono
3	Area 7 Nam..	2	functional	14	soft-good	enough	25	1639	10	groundwater	spring	gravity	gravity
4	Ezeleda	26	non functional	31	soft-good	dry	760	605	30	groundwater	shallow well	handpump	nira/tanira
5	Kwa Namaj	7	functional	28	soft-good	enough	6922	28	50	groundwater	machine dbh	submersible	submersible
6	Mission	16	non functional	9	soft-good	seasonal	103	1691	250	groundwater	shallow well	other	other

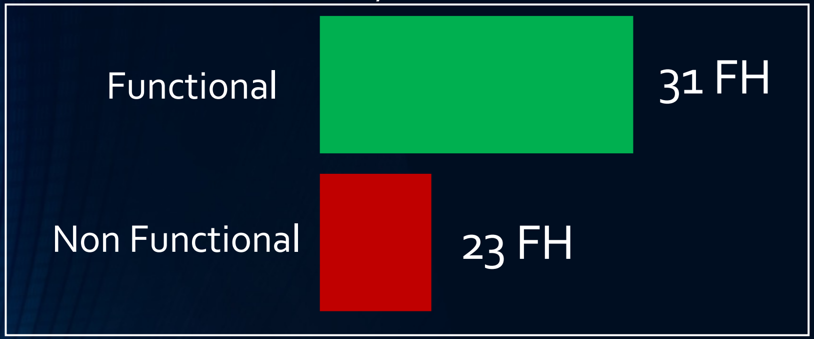
Thank you

Questions?

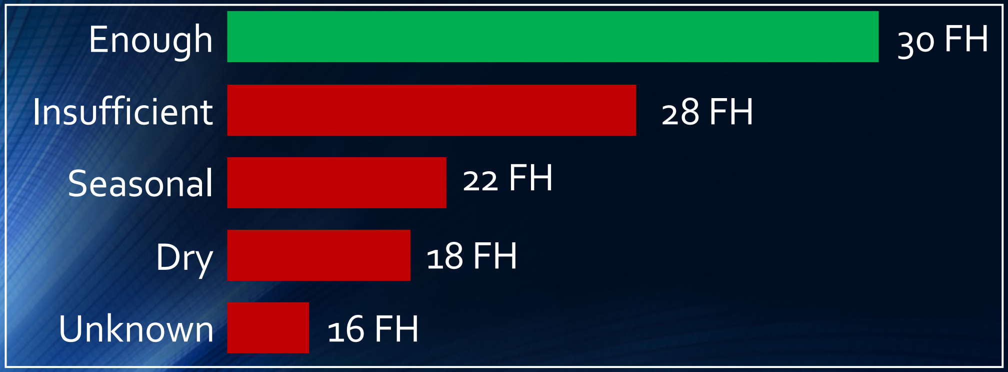
Appendix

Lower Static Head Indicates Failure

Amount of Static Head by Well Status



Amount of Static Head by Water Quantity

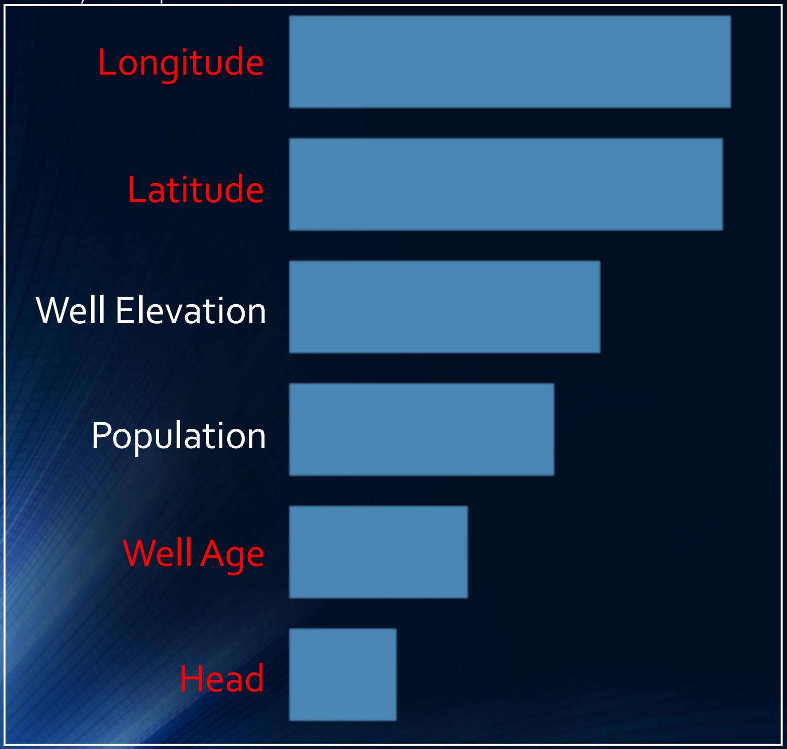


- As well ages increase, the amount of static head on the well will decrease, lowering water quantity
- This can be artificially improved by technology such as a pump
- Recommendation: Keep a close eye on static head as it directly correlates with water quantity

Most Important Features from Model

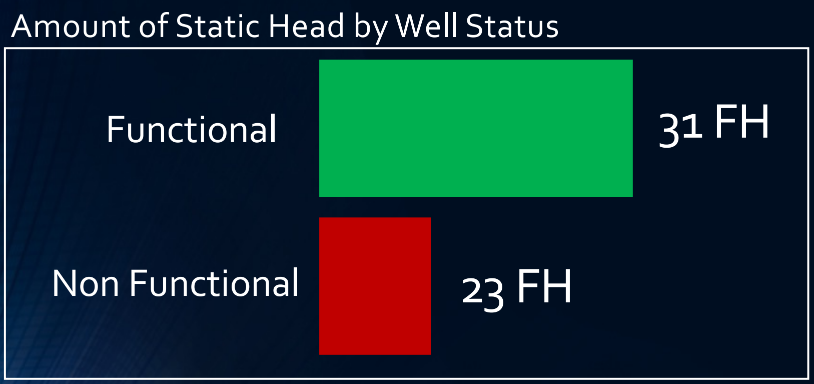
Random Forest Feature Importances

Sorted by Most Important



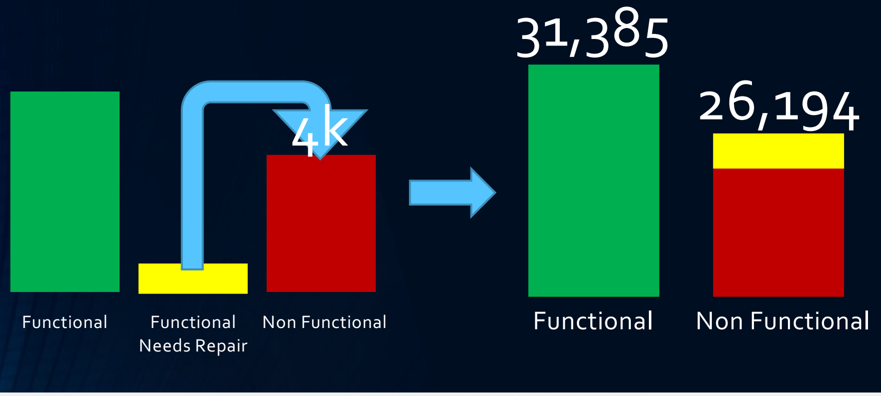
- Location (longitude and latitude) is the most important feature for predicting well function
- Well age directly affects reliability
- Well Elevation and Population are important features for prediction, although analysis did not highlight specific relationships

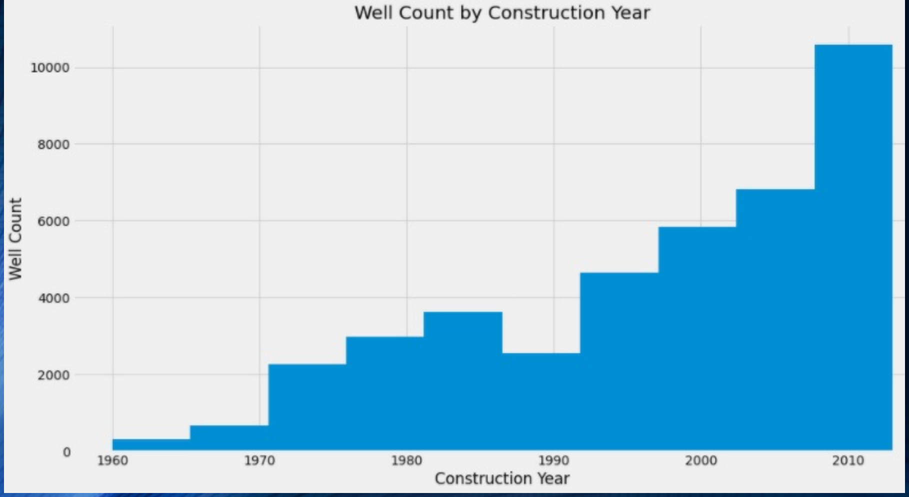
Lower Static Head Indicates Failure



- As well ages increase, the amount of static head on the well will decrease, lowering water quantity
- This can be artificially improved by technology such as a pump
- Recommendation: Keep a close eye on static head as it directly correlates with well function

Analysis Overview

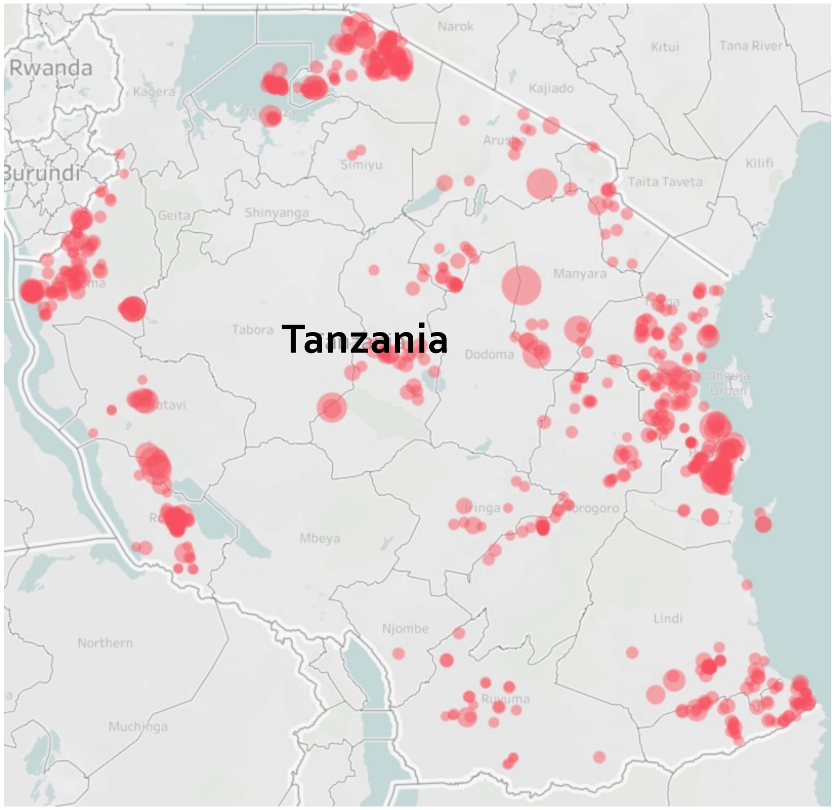




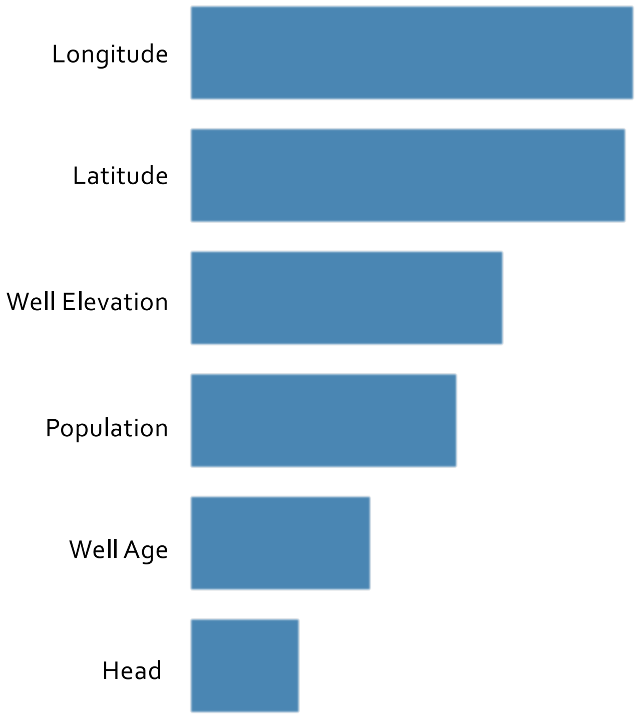
- Focused on wells not functioning or functioning at a reduced capacity (functional needs repair)
- Created well_age feature
- 20,000 wells with unknown construction_year. Filled values keeping identical distribution
- Classification modeling to use well features to predict non-functional water wells can save lives by increasing reliability and maintenance response time

Non-Functional wells that support over 1,000 citizens

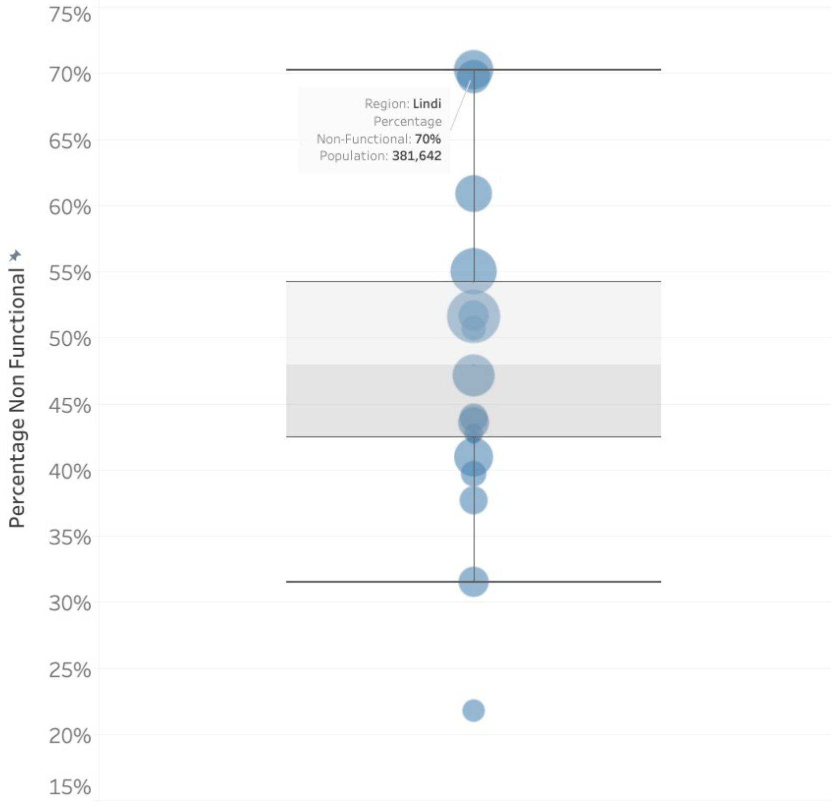
(sized by population)



Most Important Model Features

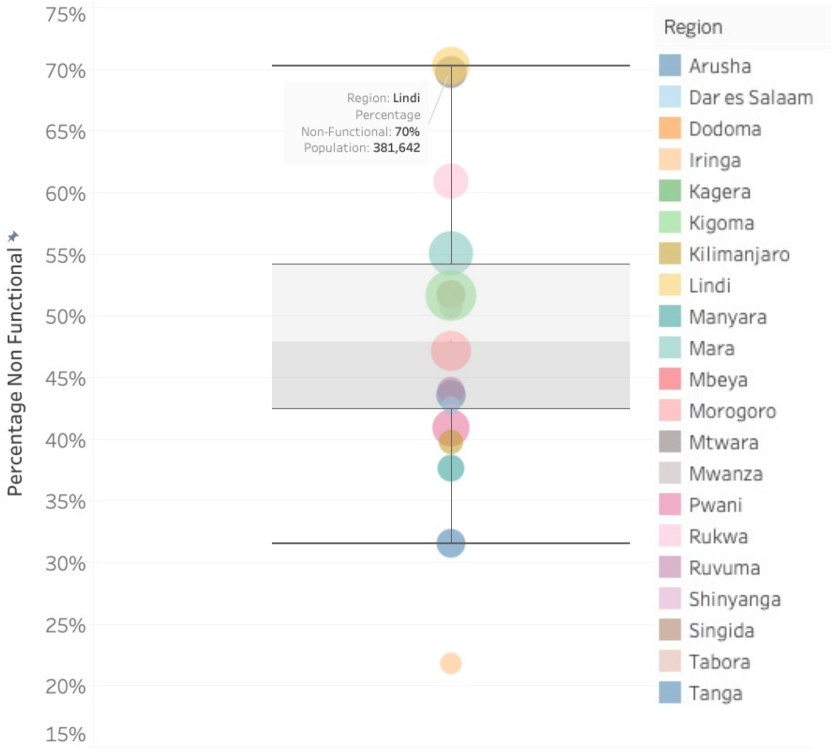


Percentage of Non-Functional Wells by Region

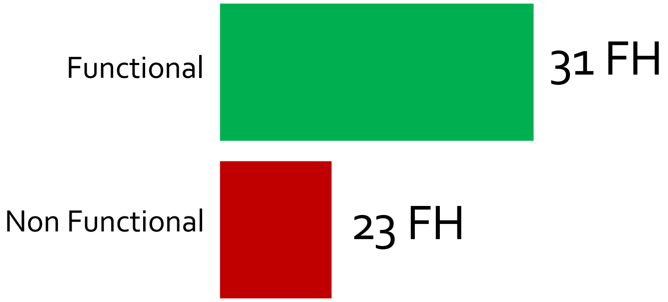


Distribution Percentage Non-Functional by Region

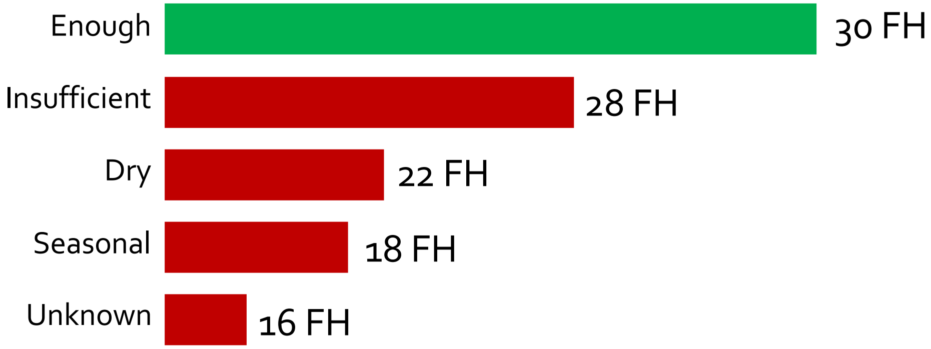
Sized by population



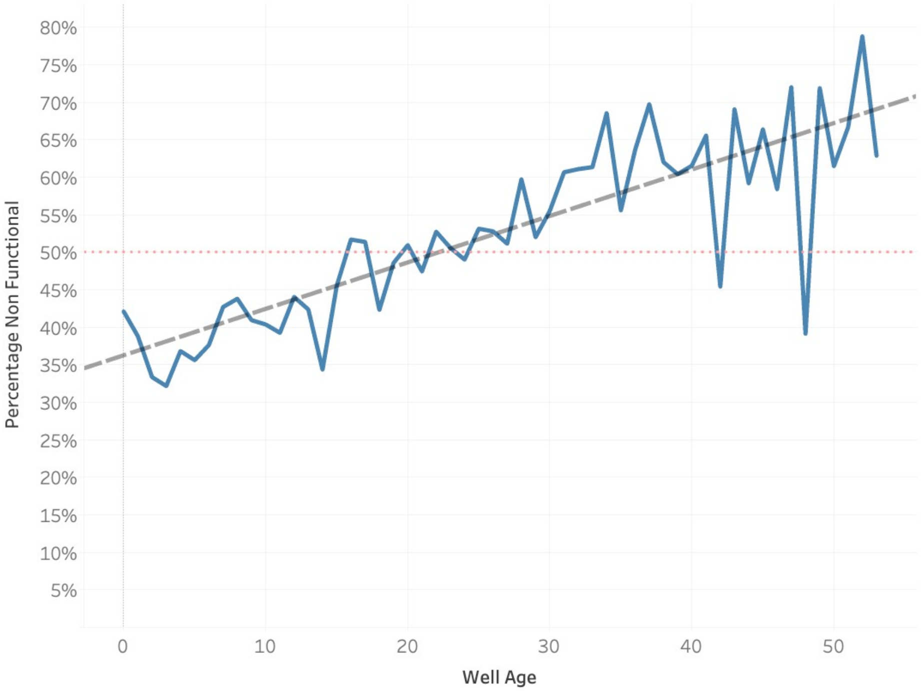
Amount of Static Head by Well Status



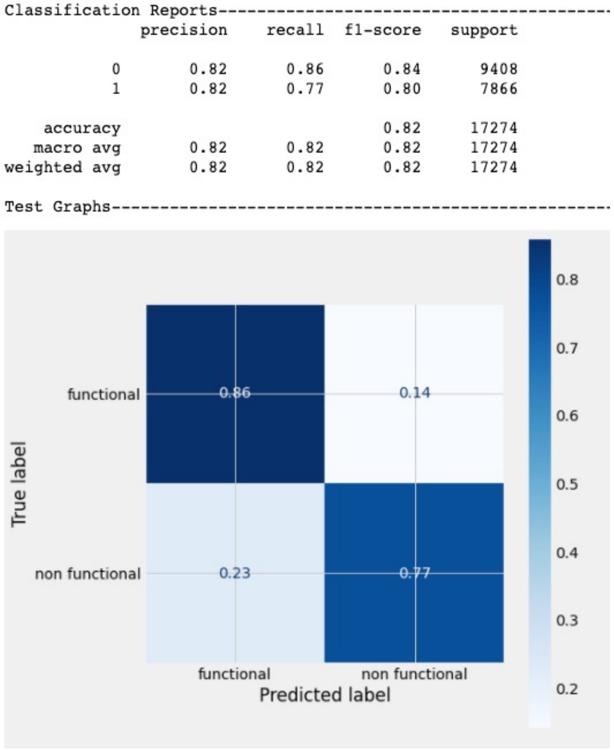
Amount of Static Head by Water Quantity



Percent Non-Functional by Well Age



Random Forest Model Results



```
Classification Reports-----------------------------------------
              precision    recall  f1-score   support

           0       0.82      0.86      0.84      9408
           1       0.82      0.77      0.80      7866

    accuracy                           0.82     17274
   macro avg       0.82      0.82      0.82     17274
weighted avg       0.82      0.82      0.82     17274

Test Graphs----------------------------------------------------
```