Water Wells in Tanzania

Using classification modeling to predict well functionality

Katarina Salcedo

Motivation - Water Crisis

- Water crisis = lack of fresh water resources to meet the standard demand
- ⅓ of the county is arid/semi-arid
- Ground and surface water are contaminated from toxic drainage systems, bacteria and human waste
- Water-borne illnesses account for over ½ of the diseases affecting nation
- Risking safety and education to walk to get water

Goal

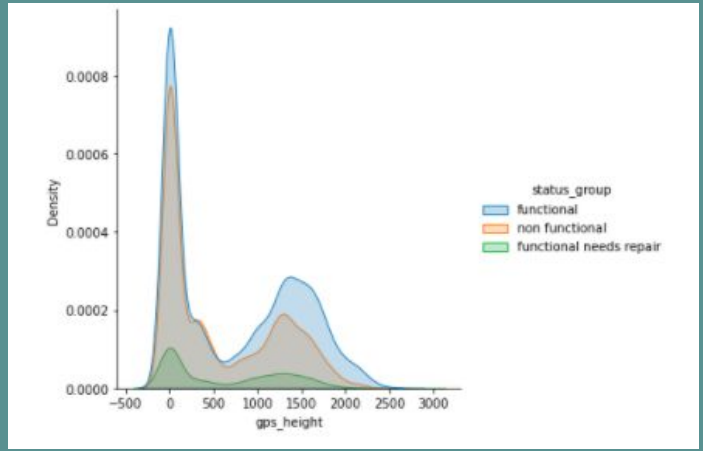
- Provide The Water Project with information on the status of wells
  - Decide where to build next
    - What type of wells to build
  - Help determine which wells need maintenance

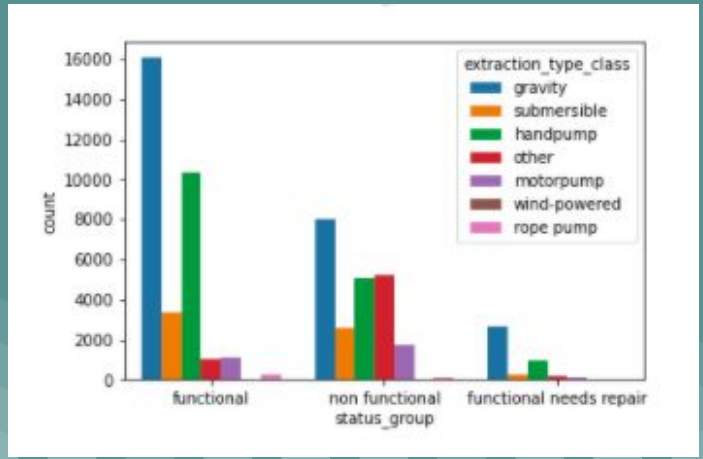
Data

- Data from Taarifa and the Tanzanian Ministry of Water
  - Contains ~59,000 rows
  - 40 independent variables that contain information about geographical location, funder, management, population, quantity and quality of water, extraction type, source, if payment is required, etc for each well
  - Target variable: status group
    - Functional
    - Non functional
    - Functional needs repair

EDA

- Looking at separability of independent variables
  - Dropped columns will little/no separability (all continuous variables)
- Dropping redundant columns i.e extraction type group, payment type, quantity group etc.
- Ended up with 16 independent variables
- Dropped null values in public meeting, scheme management, permit => ~50,000 rows
- Class imbalance:
  - 54% = functional (0)
  - 38% = non functional (1)
  - 7% = functional needs repair (2)





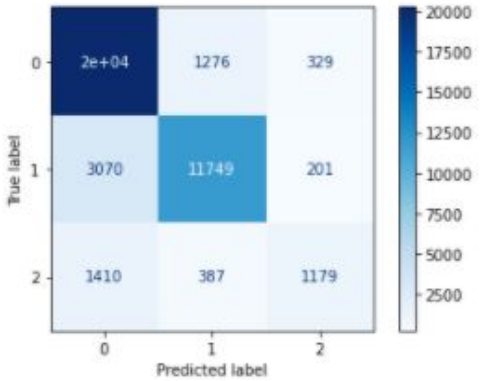
Classification report for training data:

	precision	recall	f1-score	support
0	0.82	0.93	0.87	21876
1	0.88	0.78	0.83	15020
2	0.69	0.40	0.50	2976
accuracy			0.83	39872
macro avg	0.79	0.70	0.73	39872
weighted avg	0.83	0.83	0.83	39872

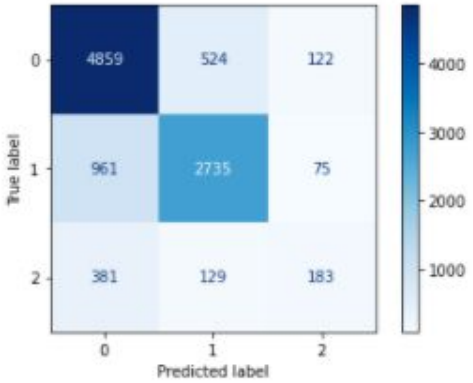
Classification report for testing data:

	precision	recall	f1-score	support
0	0.78	0.88	0.83	5505
1	0.81	0.73	0.76	3771
2	0.48	0.26	0.34	693
accuracy			0.78	9969
macro avg	0.69	0.62	0.65	9969
weighted avg	0.77	0.78	0.77	9969

Confusion Matrix Train:



Confusion Matrix Test:



Initial Model - Random Forest Classifier

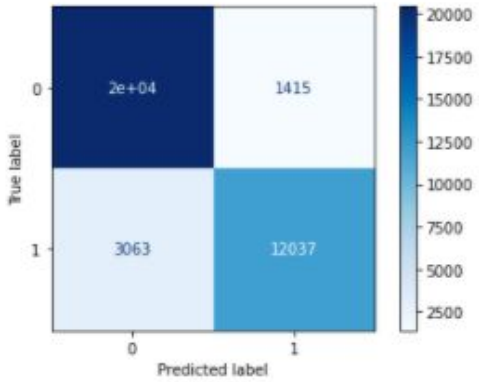
Classification report for training data:

	precision	recall	f1-score	support
0	0.87	0.94	0.90	21837
1	0.89	0.80	0.84	15100
accuracy			0.88	36937
macro avg	0.88	0.87	0.87	36937
weighted avg	0.88	0.88	0.88	36937

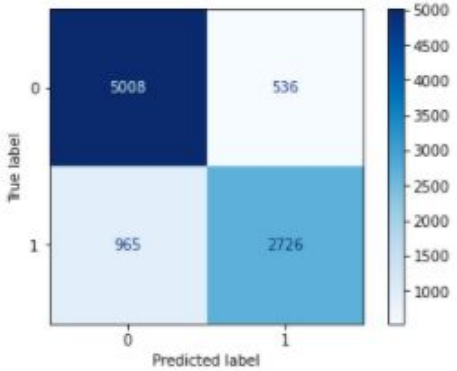
Classification report for testing data:

	precision	recall	f1-score	support
0	0.84	0.90	0.87	5544
1	0.84	0.74	0.78	3691
accuracy			0.84	9235
macro avg	0.84	0.82	0.83	9235
weighted avg	0.84	0.84	0.84	9235

Confusion Matrix Train:



Confusion Matrix Test:



Final Model - Random Forest

Conclusions

- This model is able to predict the functionality of a water well with 84% accuracy.
- Most important features in determining this are:
  - Quantity
  - Waterpoint type
  - Extraction type
  - Payment
- Next steps:
  - Improving accuracy
  - More feature engineering i.e getting the age of a well
  - Plotting locations of nonfunctional wells to see trends - identify areas with greater need

THANK YOU

Email: ksalcedo04@gmail.com

GitHub repo