Water in Tanzania

Predicting Conditions of Water Wells with a Machine Learning Classifier

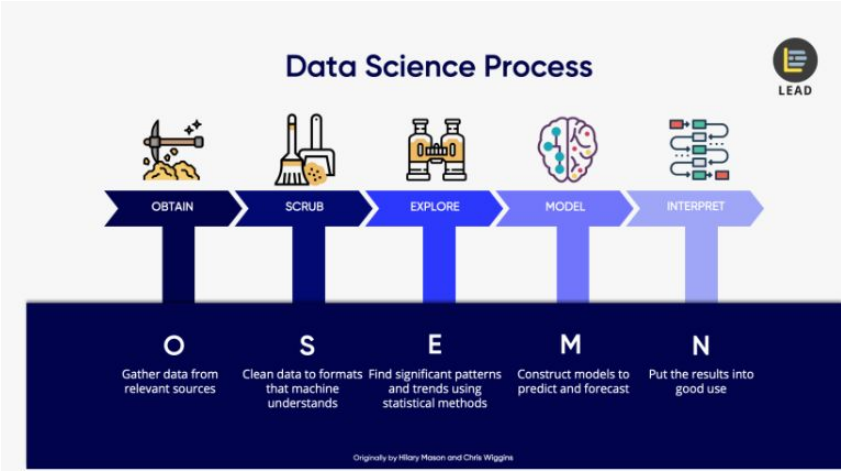
Kyaw Saw Htoon

Problem Statement

- Availability of clean and potable water is vital
- Important to ensure that water wells are working
- Develop a Machine Learning Classifier to predict the conditions of water wells
  - Functional
  - Functional but needs repair
  - Non-Functional
- Helps to improve maintenance operations

Model Development

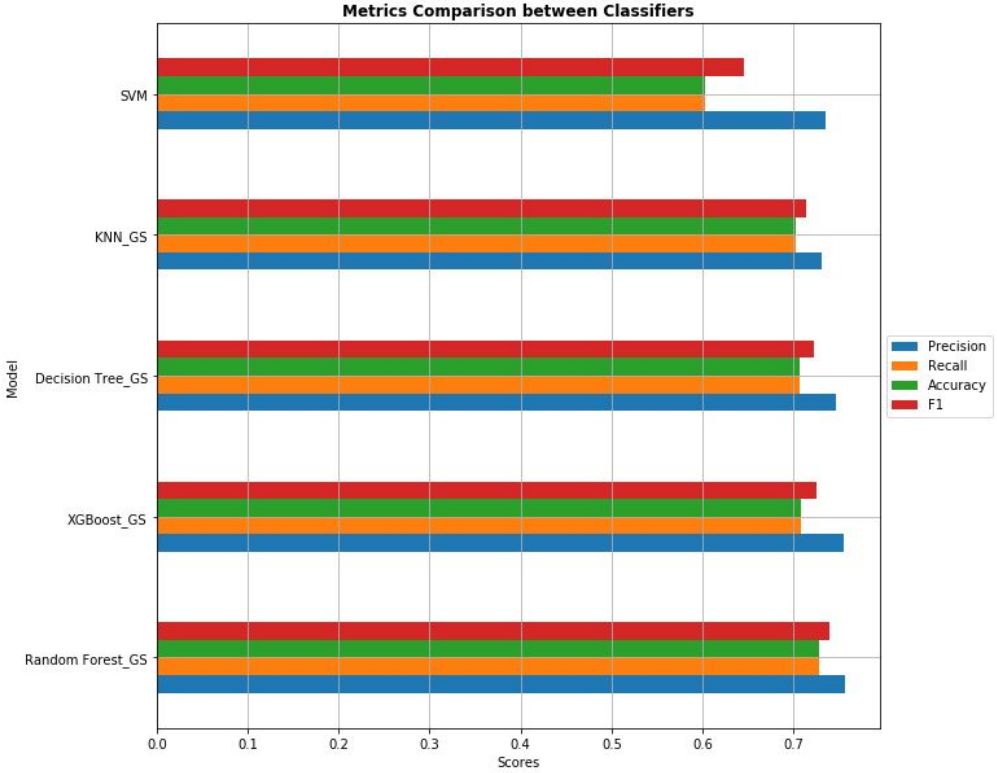
- Original data is provided by Taarifa and Tanzanian Ministry of Water
- 39 estimators to predict the conditions of water wells
- Follow OSEMN Framework
    - Obtain
    - Scrub
    - Explore
    - Model
    - Interpret



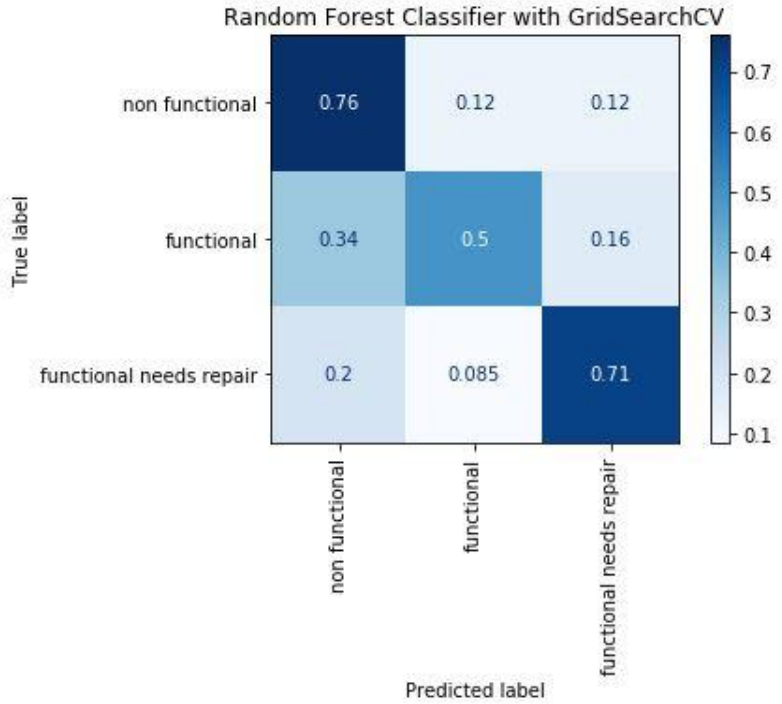
Model Selection

- Developed 5 different Classifiers
- Based on Accuracy and F1 scores

Model	Precision	Recall	Accuracy	F1
Random Forest_GS	0.757020	0.728754	0.728754	0.740070
XGBoost_GS	0.755069	0.707811	0.707811	0.724816
Decision Tree_GS	0.747335	0.706801	0.706801	0.722288
KNN_GS	0.730554	0.703165	0.703165	0.713880
SVM	0.735454	0.603367	0.603367	0.645695



Random Forest Classifier



Training Data

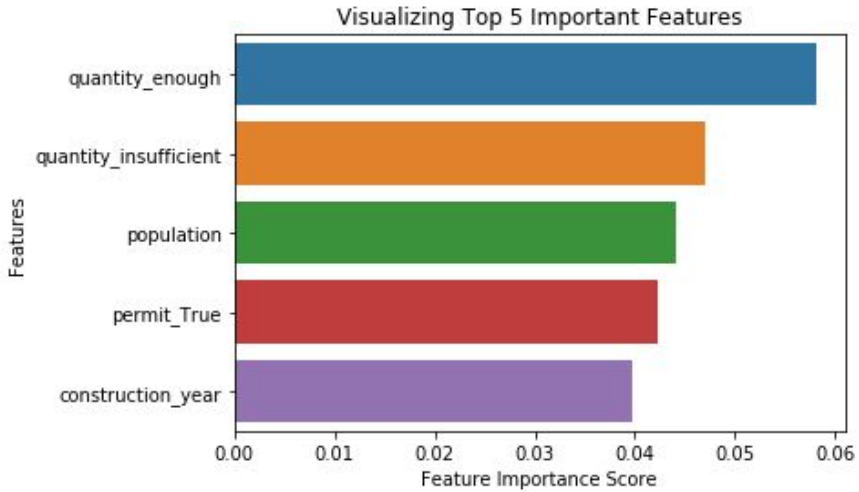
	precision	recall	f1-score	support
functional	0.81	0.81	0.81	24161
functional needs repair	0.83	0.90	0.87	24161
non functional	0.89	0.82	0.86	24161
accuracy			0.84	72483
macro avg	0.85	0.84	0.84	72483
weighted avg	0.85	0.84	0.84	72483

Testing Data

	precision	recall	f1-score	support
functional	0.80	0.76	0.78	8098
functional needs repair	0.28	0.49	0.36	1074
non functional	0.78	0.72	0.75	5678
accuracy			0.73	14850
macro avg	0.62	0.66	0.63	14850
weighted avg	0.76	0.73	0.74	14850

Important Estimators

- 81 estimators including dummies



- Impurity based feature importance

Areas of Further Improvement

- Get a better understanding of estimators
- Use a broader range of hyperparameter inputs
- Construct the classifier with all the original features
- User more complex ensemble models (e.g. Stacking)

The End

Thank you