Tanzanian Water Well Data

By Keene Kelderman

Background on the data

Processing and Choosing Predictor Variables

```
Values for column: extraction_type
Number of factors: 18
gravity                        26780
nira/tanira                     8154
other                           6430
submersible                     4764
swn 80                          3670
mono                            2865
india mark ii                   2400
afridev                         1770
ksb                             1415
other - rope pump                451
other - swn 81                   229
windmill                         117
india mark iii                    98
cemo                              90
other - play pump                 85
walimi                            48
climax                            32
other - mkulima/shinyanga          2
Name: extraction_type, dtype: int64
```

```
Values for column: extraction_type_group
Number of factors: 13
gravity              26780
nira/tanira           8154
other                 6430
submersible           6179
swn 80                3670
mono                  2865
india mark ii         2400
afridev               1770
rope pump              451
other handpump         364
other motorpump        122
wind-powered           117
india mark iii          98
Name: extraction_type_group, dtype: int64
```

```
Values for column: extraction_type_class
Number of factors: 7
gravity          26780
handpump         16456
other             6430
submersible       6179
motorpump         2987
rope pump          451
wind-powered       117
Name: extraction_type_class, dtype: int64
```

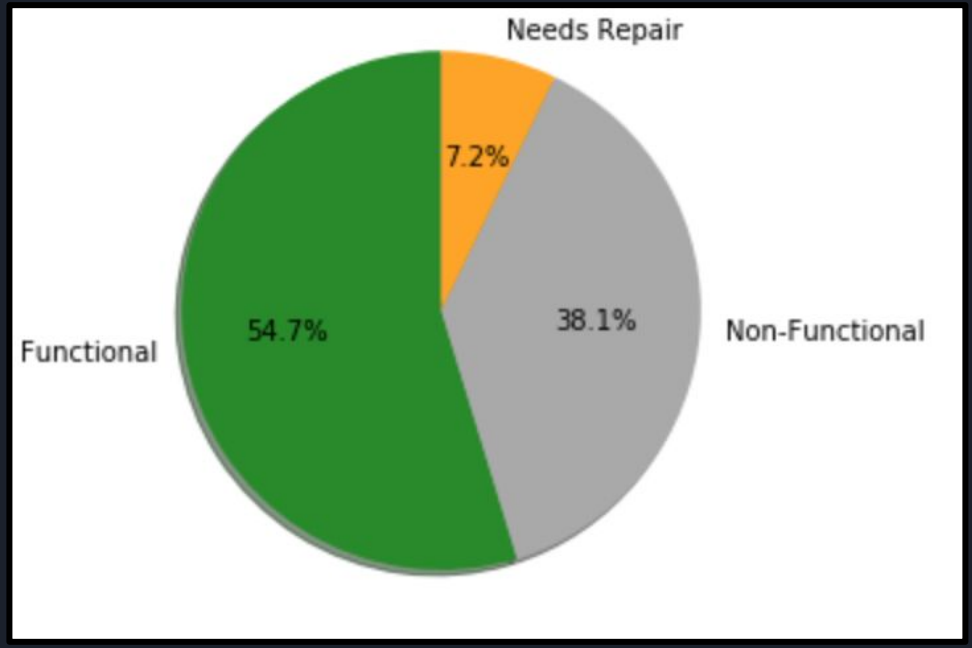
Classifier Methods Used

Models:

1) Decision Tree
2) Random Forest Classifier
3) Logistic Regression
4) XGBoost

Sampling Technique:

1) Raw, unbalanced data
2) Up-sampled data
3) SMOTE



Model Results

	non_func_prec	non_func_rec	func_prec	func_rec	repair_prec	repair_rec	accuracy
d_tree_norm	0.751	0.760	0.800	0.800	0.402	0.375	0.755
d_tree_up	0.756	0.729	0.795	0.768	0.321	0.472	0.733
d_tree_smote	0.740	0.755	0.808	0.778	0.340	0.402	0.743
r_forest_norm	0.809	0.766	0.803	0.860	0.488	0.355	0.789
r_forest_up	0.804	0.752	0.810	0.806	0.347	0.480	0.763
r_forest_smote	0.792	0.778	0.817	0.824	0.414	0.424	0.779
log_reg_norm	0.774	0.611	0.710	0.902	0.727	0.009	0.729
log_reg_up	0.779	0.610	0.772	0.636	0.168	0.602	0.624
log_reg_smote	0.761	0.585	0.738	0.806	0.211	0.324	0.689
xgb_norm	0.834	0.613	0.718	0.934	0.687	0.051	0.750
xgb_up	0.818	0.603	0.784	0.689	0.201	0.681	0.656
xgb_smote	0.795	0.605	0.765	0.789	0.230	0.470	0.697

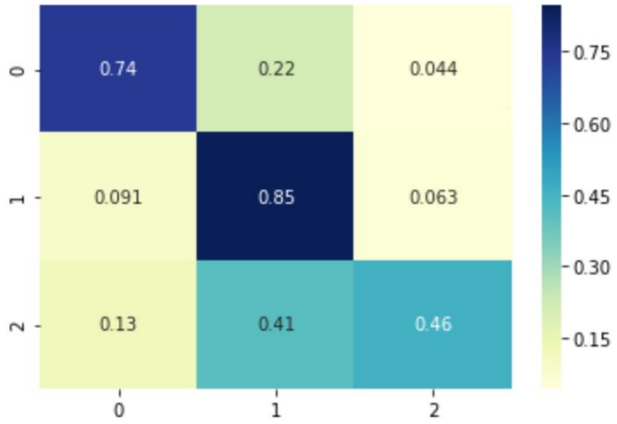
Model Results: Final Parameters

- Criterion : ‘gini’
- Max_depth: None
- Min_samples_leaf: 3
- Min_samples_split: 5
- N_estimators: 100

Classification Report

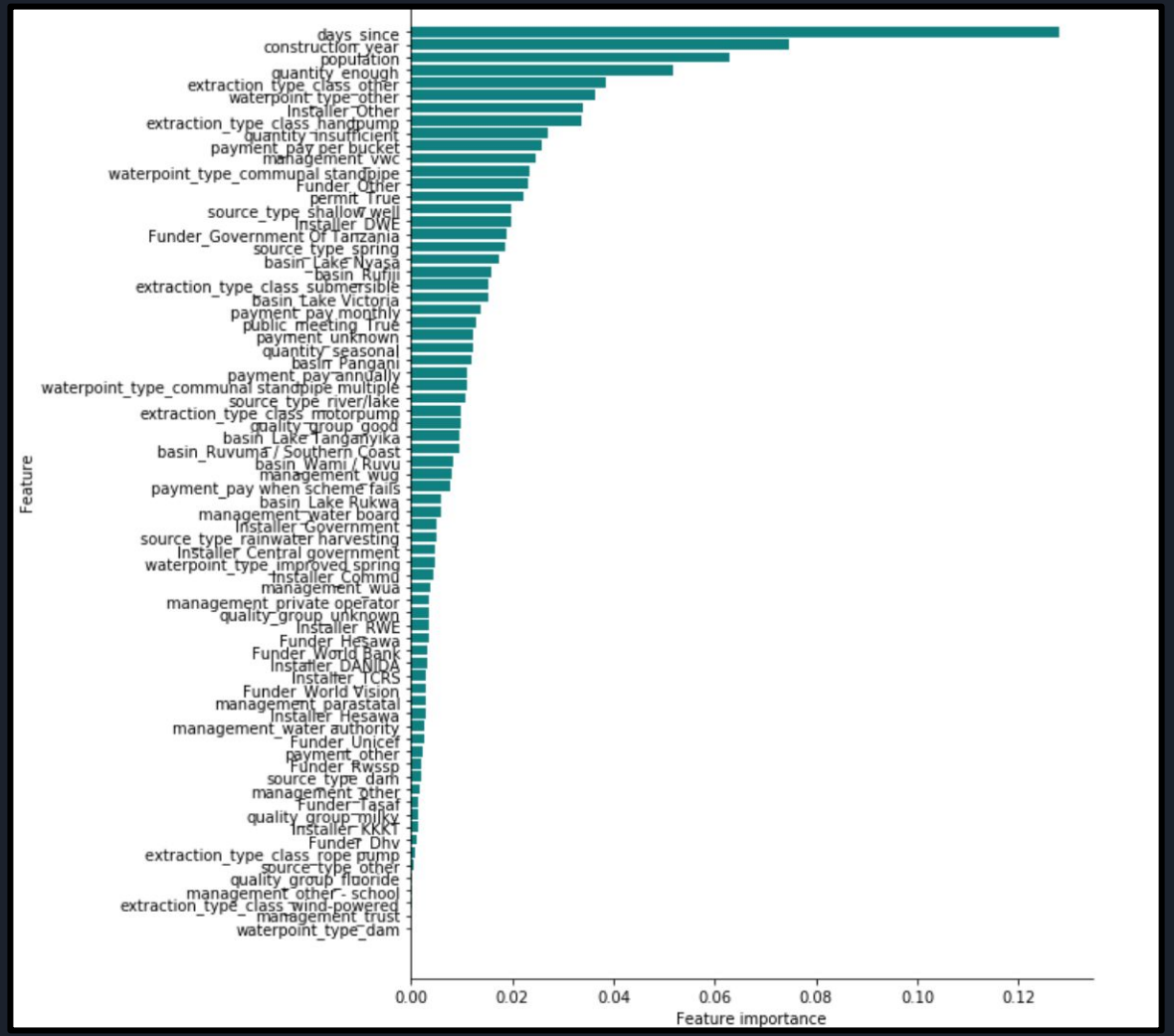
	precision	recall	f1-score	support
0	0.826	0.740	0.781	4880
1	0.808	0.846	0.827	7088
2	0.385	0.460	0.419	897
accuracy			0.779	12865
macro avg	0.673	0.682	0.675	12865
weighted avg	0.785	0.779	0.781	12865

Confusion Matrix



Interpretation of final model

1) Days Since
2) Construction Year
3) Population
4) Quantity_Enough
5) Extraction_Type_Class_Other



Further Research

1) XGBoost Model Using Up-sampled training data

2) Data Reclassification

Thanks for listening!