FUNCTIONALITY OF TANZANIAN WATER WELLS

Matthew Andrews

OVERVIEW

Driven Data competition>	Pump it up: Data Mining the Water Table	-2015
	Reopened and ongoing	>6 more months
Taarifa:	Winner of the London Water Hackathon	-2011
	Deployed in Tanzania	-2012

Prediction of functionality -for better use of resources

DATA

59400 Wells

39 Features:

- Location
- Logistics
- Pump
- Water

Many similar hierarchical features, e.g.:
extraction_type_class < extraction_type_group < extraction_type

Class Distribution

Functional:	54%
Need's repair:	7%
Non-Functional:	38%

EXPLORATORY DATA ANALYSIS

Categorical Data:

30 out of 39 features

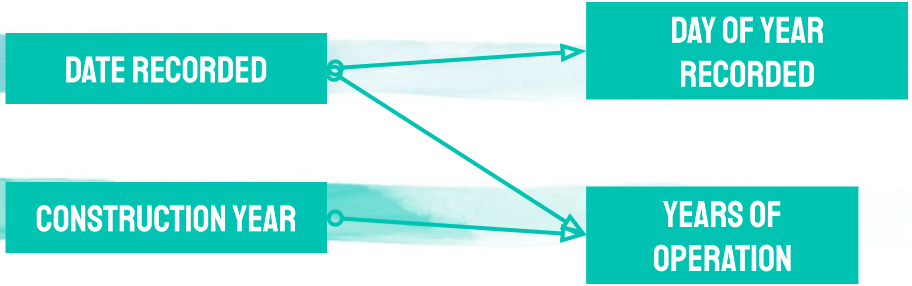
Some with thousands of categories

Dropped Features
Trimmed Categories
One Hot Encoded

Dropped Features
Target encoding

Catboost

EXPLORATORY DATA ANALYSIS



DATE RECORDED

CONSTRUCTION YEAR

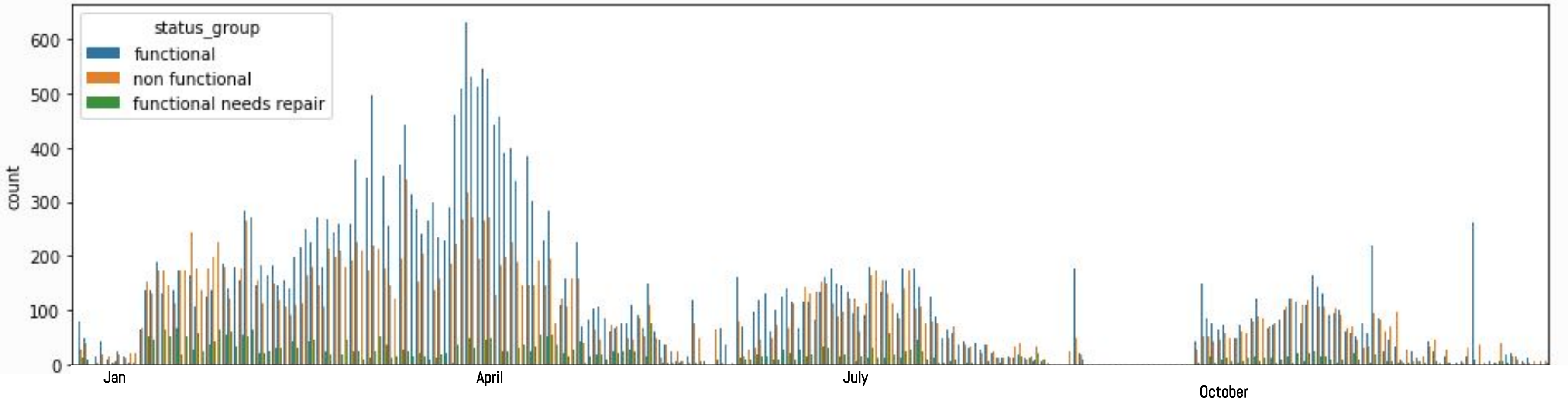
DAY OF YEAR RECORDED

YEARS OF OPERATION

Climate:

'Long Rains': March, April, May

Higher ratio of well pumps found functional during the rainy season.

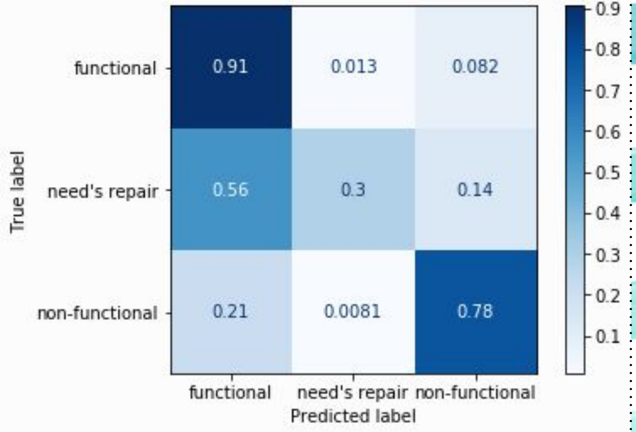


MODELS

Model	Accuracy
Logistic Regression	0.71
K Nearest Neighbors	0.78
Naive Bayes	0.69
Decision Tree	0.78
Random Forest	0.73
Adaboost	0.74
Gradient Boosting	0.80
Support Vector Classifier	0.80
XGBoost	0.74
CatBoost	0.81

Model: CatBoost

- Highest Accuracy
- Fastest



Poor predictor of ‘need's repair’ -minority class(7%)

- mostly predicted as ‘functional’

CONCLUSION

CatBoost model -81% accurate

Important features: mostly location based

- Same source?
- Recorded at the same time?

More data:

- Seasonally functional -as a class
- When did a well pump become non-functional

FUTURE WORK

- SMOTE on the smallest class
- More hyperparameter tuning in XGBoost and SVC
- Model bagging

- Use CatBoost for imputing
- Try H2O algorithm

THANK YOU

Thanks also to:

- Taarifa
- Tanzanian Ministry of Water
- DrivenData

Matthew Andrews

GitHub Repo:

https://github.com/Maltanno/Phase3_Project