Pump it Up: Data Mining the Water Table

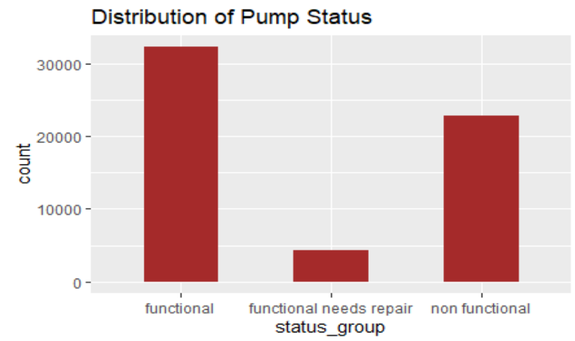
Team Members: Nisha Rangnani, Jaishree Palaniswamy, Nishitha Maniganahalli

IST707- DATA ANALYTICS

Prof. Ying Lin

Problem Statement

This dataset consists of details about the pumps that are available in Tanzania. A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania. In order to find the faulty pump, we are going to predict the condition of the pump(functional, non-functional, functional needs repair) based on a number of variables about what kind of pump is operating, when it was installed, and how it is managed.



Data Description

Our dataset consists of 59400 rows and 41 columns with the details of pump locality, water extraction type, quality, quantity of water, source from where the pump gets the water and many more.

Models Used

To predict the status of the pump operations, we are using different classification models.

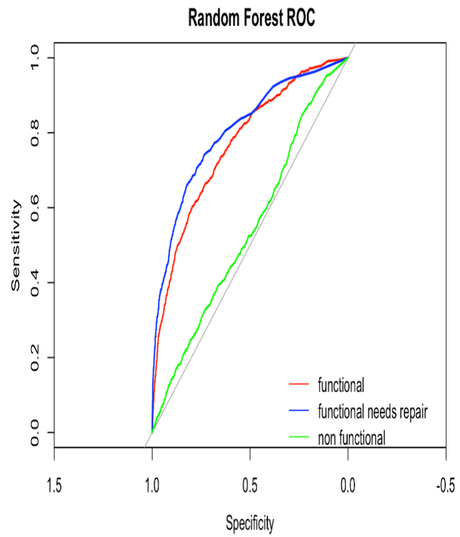
MODEL	ACCURACY (IN %)
Decision Tree	75.56
Random Forest Classifier	78.40
Gradient Boosting Machines	73.88
Support Vector Machine	75.0
Extreme Gradient Boosting	75.49
Association Rule Mining	

Model Comparison:

We tested Decision Tree, Random Forest Classifier, Gradient Boosting Machine, Support Vector Machine and Extreme Gradient Boosting to predict the faulty pumps. Here, we have compared the best two models which has higher accuracy.

Decision Tree

Random Forest



Performance Evaluation

Training and testing data were split into 80-20 split and below are the results of two models in form of confusion matrix.

Decision Tree

Predicted

Actual	F	FNR	NF
F	5493	470	1242
FNR	166	241	81
NF	792	152	3241

Random Forest

	F	FNR	NF
F	5624	454	1101
FNR	190	277	82
NF	637	132	3381

F – Functional **FNR – Functional needs repair**

NF – Non functional

Conclusion

Based on various evaluation metrics for different tasks, we conclude that Random forest performs the best faulty pump prediction with 78.40% accuracy.

Algorithm	Accuracy	AUC Score	Time taken
Decision Tree	83.7%		~1800 sec
Random Forest	83.01%	77.20	~10 sec

Shiny App Link

The link provided below will allow the user to modify various parameters in order to predict the condition of the pump.