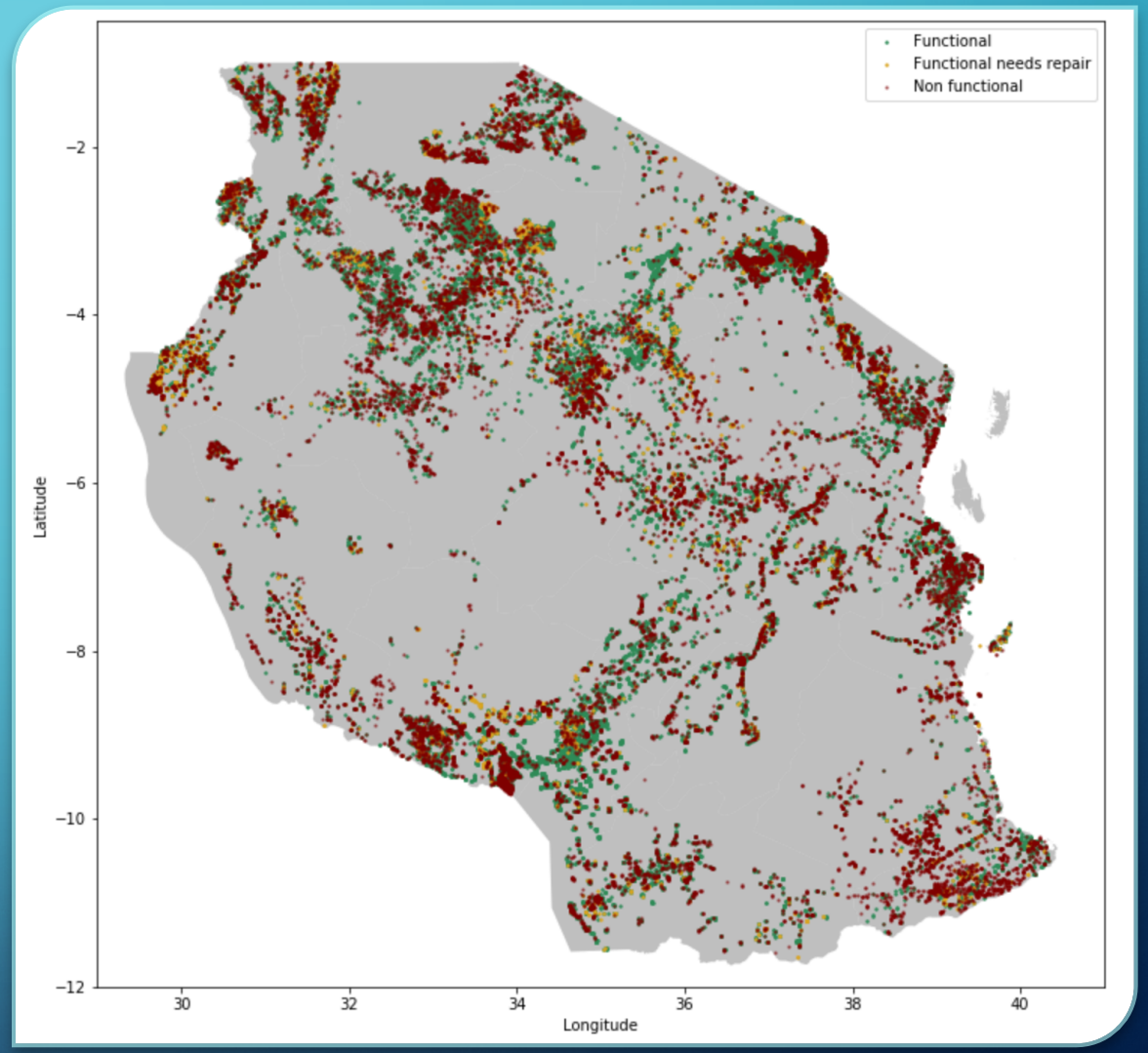
MOD 3 PROJECT

ANILA QURESHI & LUC BATTY

PROJECT OVERVIEW: TANZANIAN WATER WELLS

- The goal of this project was to create a model capable of predicting the functioning status of water wells in Tanzania.
- The dataset included approximately 60’000 individual wells located across Tanzania, each classified as either functional, functional but needing repairs, and non-functional.



FUNDAMENTAL PROJECT QUESTIONS

- What kind of results are most useful to the stakeholders?
- How strongly can we predict the status of a well?
- What are the most important predictor variables?

STAKEHOLDER ANALYSIS

Responsible parties

User communities

Both are primarily interested in identifying non-functional wells and wells in need of repair. Time taken to identify and repair non-functional wells, as well as travel costs to remote locations would ideally be minimized. Therefore the model should aim to minimize false positives.

MODEL RESULTS (BINARY)

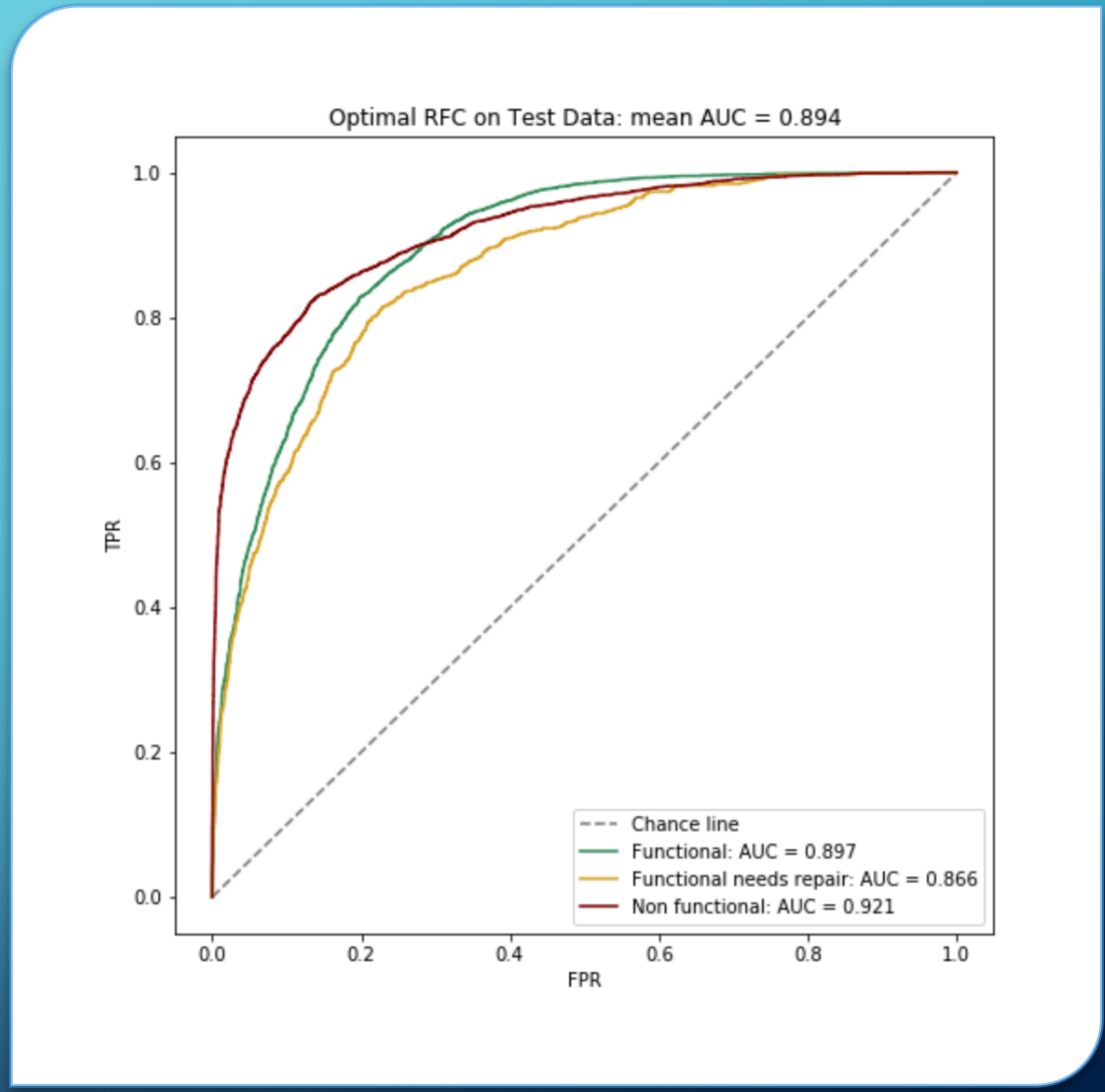
Model	Accuracy	AUC
LogReg (Baseline)	0.71	0.786
LogReg	0.75	0.824

MODEL RESULTS (TERNARY)

Model	Accuracy	Bal. Accuracy	AUC_f	AUC_fnr	AUC_nf	AUC_mean
DTC (Baseline)	0.744	0.629	0.774	0.651	0.805	0.744
RFC	0.805	0.643	0.897	0.866	0.921	0.894
ABC	0.732	0.518	0.802	0.759	0.827	0.796

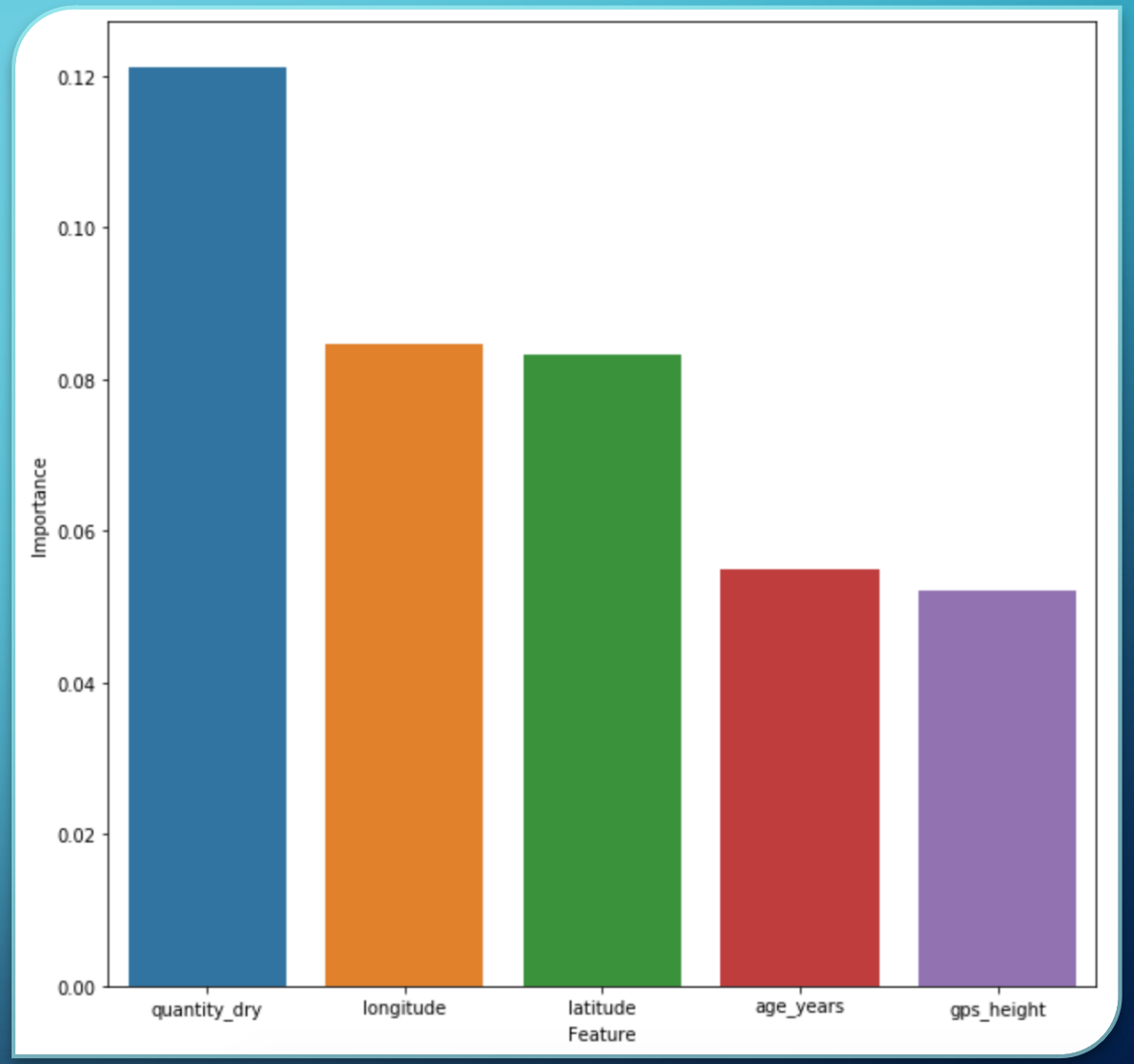
BEST OVERALL MODEL

THE BEST-PERFORMING MODEL OVERALL WAS THE OPTIMIZED TERNARY RANDOM FOREST CLASSIFIER



MOST IMPORTANT FEATURES

- Using the optimal model, the top five predictive features were identified as:
- Dry source
- Longitude
- Latitude
- Age in years
- Altitude



FUTURE WORK

Run additional models (e.g., KNN, XGBC) to determine if the predictive power could be improved upon

Carry out sector research in order to estimate the economic and social costs of false positives and false negatives

Carry out additional research on the impact of the features identified as the most important, and their effect on well functionality