

MACHINE LEARNING AND ANALYSIS: DATA SCIENCE

TANZANIAN WATER PUMP PROJECT

INTRODUCTION

INTRODUCTION

- ▶ 1/6 of the world population lack access to safe water;
- ▶ The average African uses 5 gallon of water daily;
- ▶ How can we predict whether a Tanzanian water pump is functional or not?
- ▶ **Relevance:** overall implication to public policy, governmental agencies, NGOs, general public.

DATA AND ANALYSIS

DATA UNDERSTANDING

- ▶ Databases

- ▶ Taarifa and Tanzanian Ministry of Water (from DrivenData);
 - ▶ Construction years from 1960-2013, with 59,400 rows;

- ▶ Variables

- ▶ Status, location, extraction type, source, quantity, construction year, management, etc.

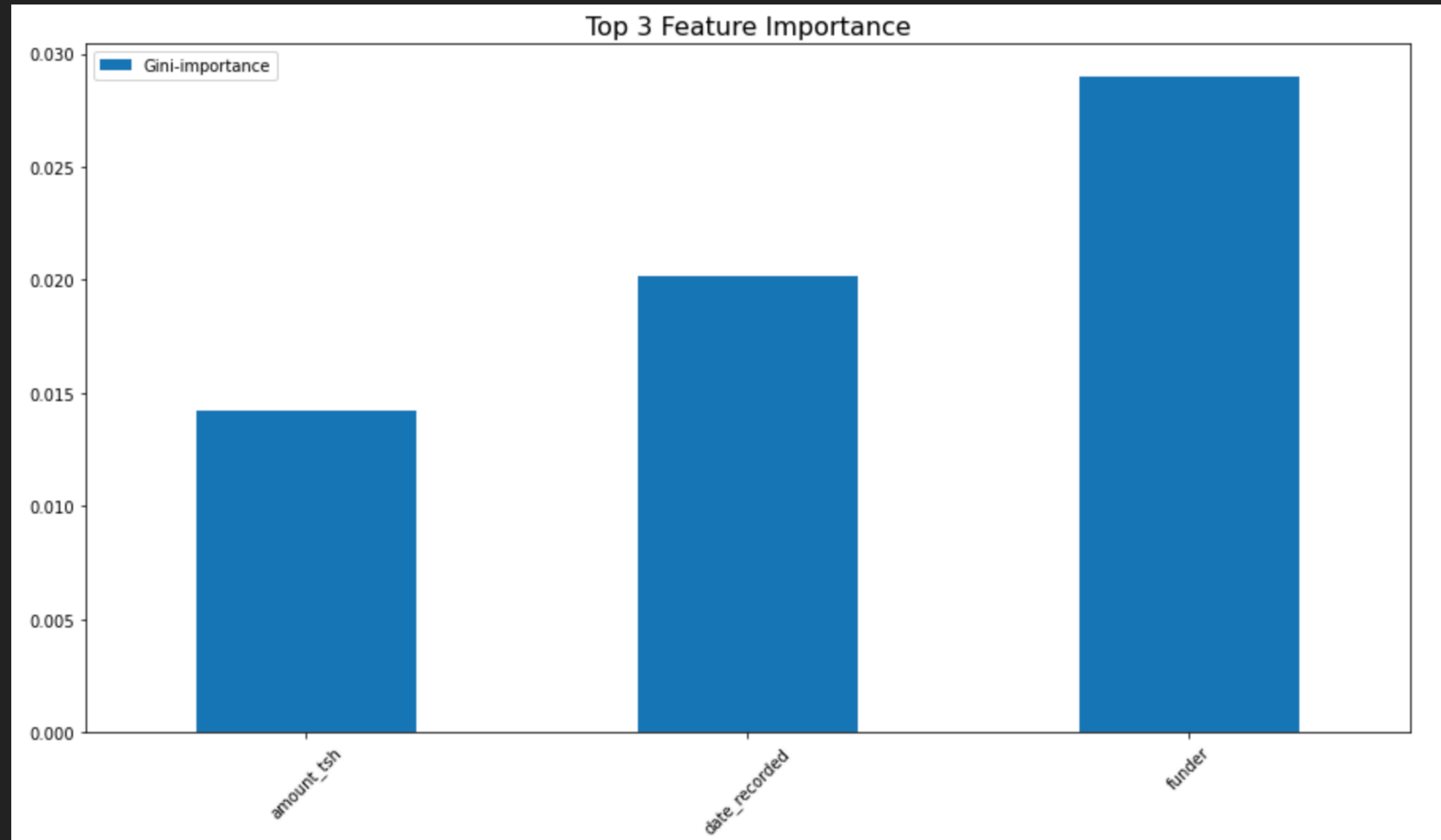
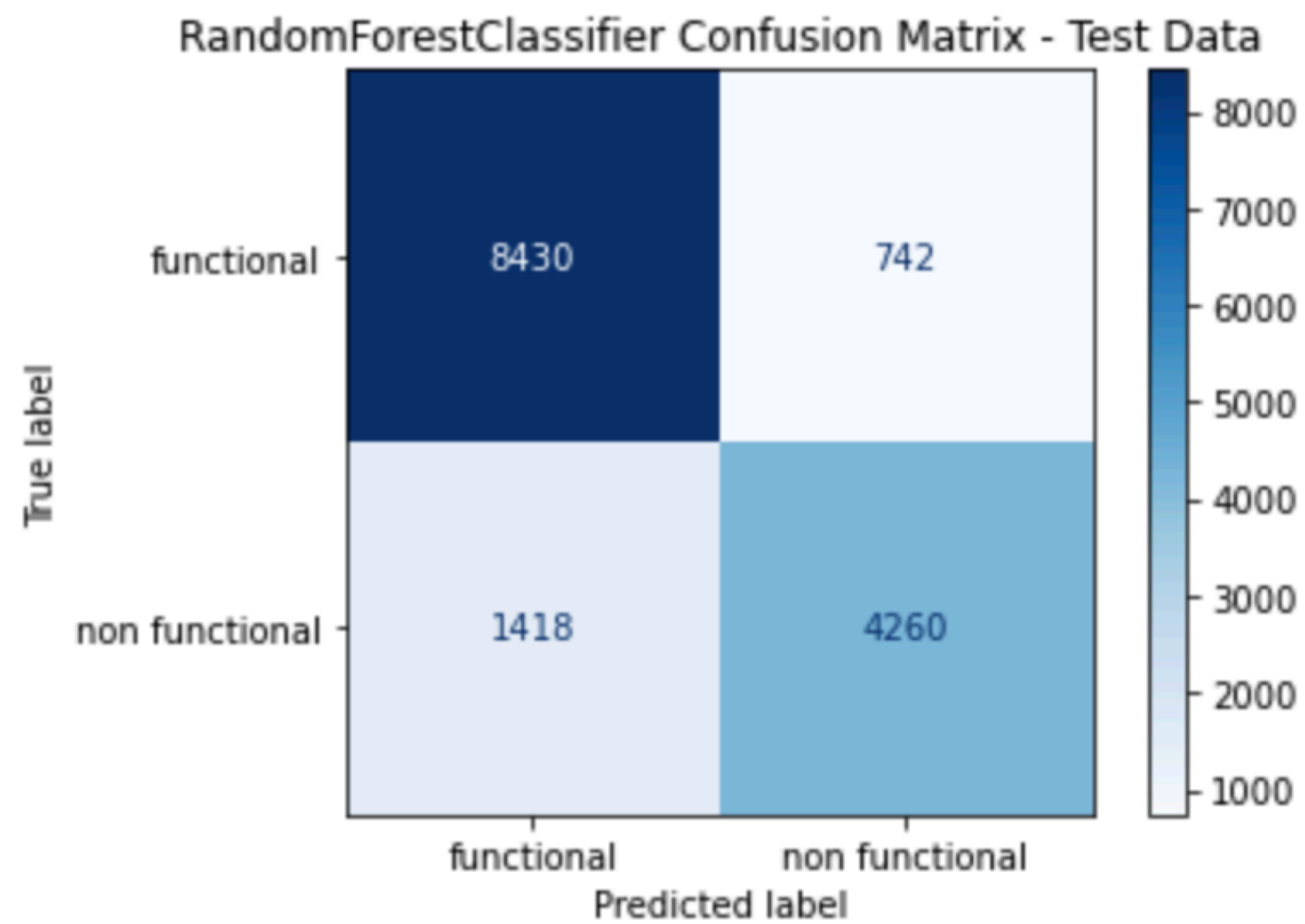
- ▶ Additional feature

- ▶ Geo-plotting of water pumps by status.

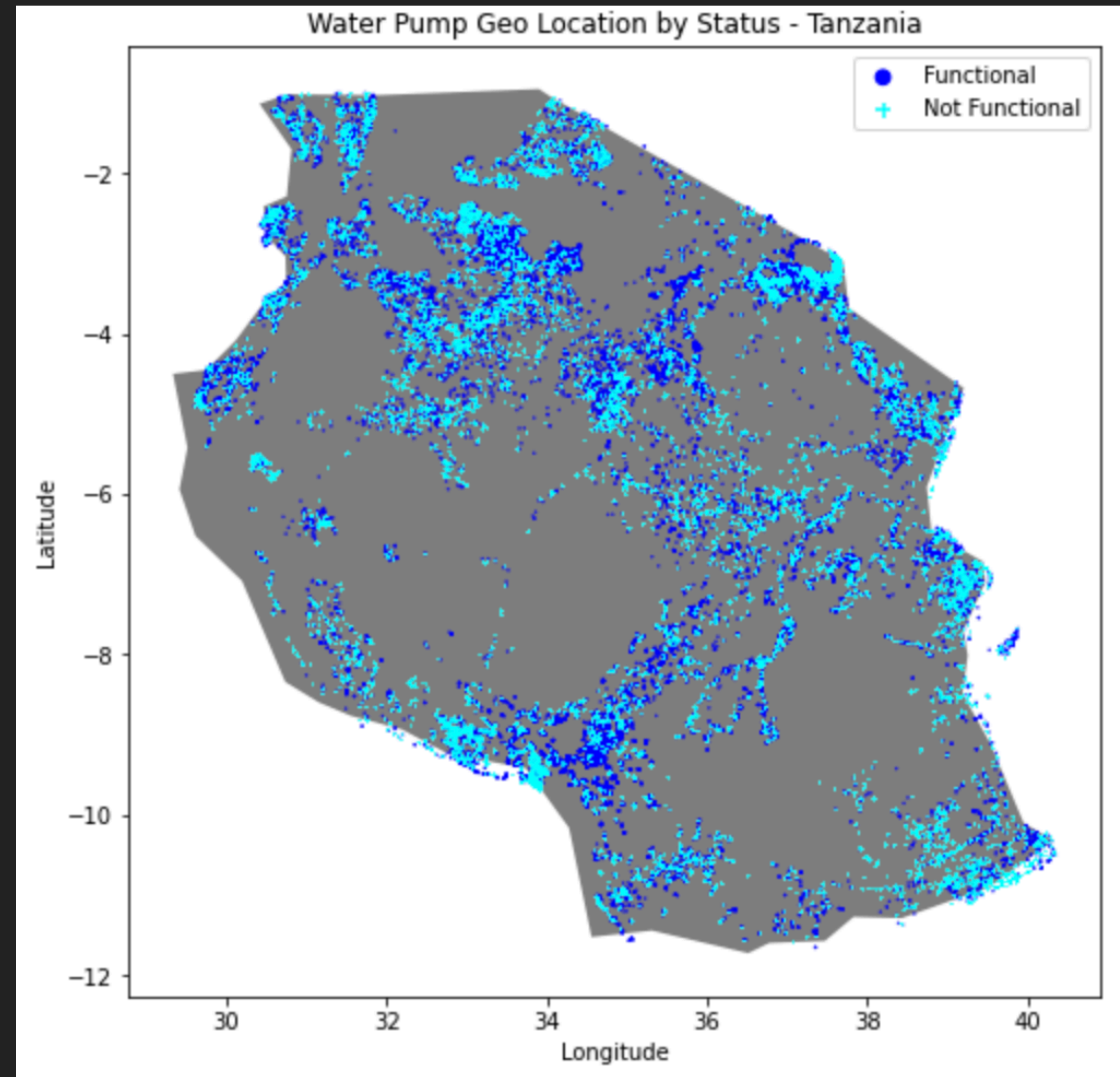
MACHINE LEARNING

- ▶ Three models
 - ▶ logistic regression (baseline), decision tree and random forest;
- ▶ Key metric
 - ▶ Balance of false positives and false negatives (f1-score);
 - ▶ RandomForestClassifier has a 85% accuracy and weighted f1-score;
- ▶ Feature Importances
 - ▶ Funder, date recorded and amount of water available.

CONFUSION MATRIX AND FEATURE IMPORTANCE



TANZANIAN WATER PUMP BY STATUS



CLOSING REMARKS

STRATEGIES

- ▶ **Prophylactic:** the best model is able to predict whether a pump is working or not by 90%;
 - ▶ This can lead to better planning on when to fix functional pumps;
 - ▶ Funder, date recorded and water amount are the strongest predictors;
- ▶ **Expansion:** areas with less pumps can be used for expansion and diminishing traveling distance/time.

LIMITATIONS

- ▶ Adding demographic data about each specific area;
- ▶ Predicting pumps that are functional and need repair;
- ▶ Matching with more robust numerical data can improve the models;
 - ▶ The data is noisy, discretion and refinement is advised.

Thank you!

<https://github.com/ovilar>