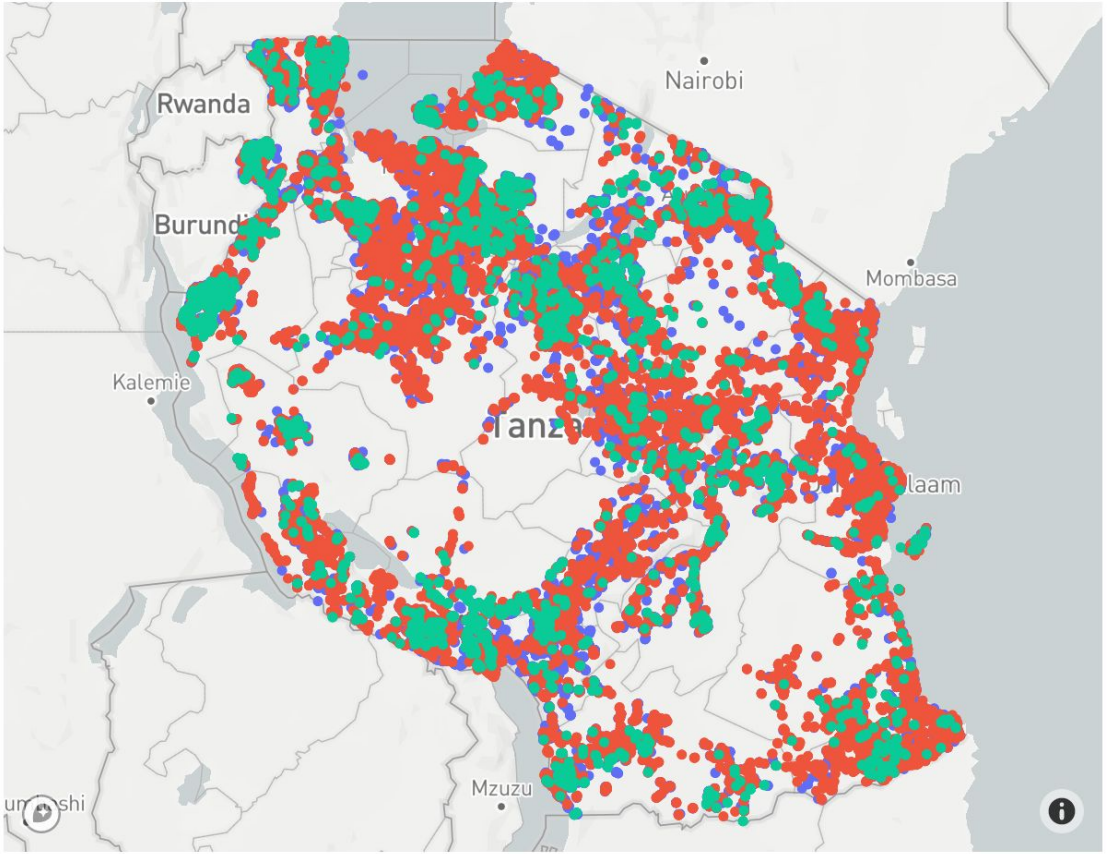
Tanzanian Water Well Data

David Cuervo

Background

- Competition through DrivenData
- Data collected from Taarifa and the Tanzanian Ministry of Water
- Predicting well functionality based on a number of variables about the well’s age, location, and how it is managed



- status_group=functional
- status_group=non functional
- status_group=functional needs repair

Data Exploration and Cleaning

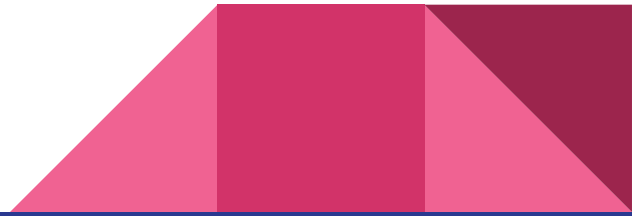
- Went through each variable to deal with missing data and outliers
- Dropped columns
- Made dummy variables to replace the categorical variables in the dataset

Columns Dropped

- date_recorded
- wpt_name
- num_private
- region_code
- district_code
- public_meeting
- recorded_by
- scheme_name
- extraction_type
- extraction_type_class
- Management
- management_group
- Payment
- water_quality
- quantity_group
- Source
- source_class
- waterpoint_type

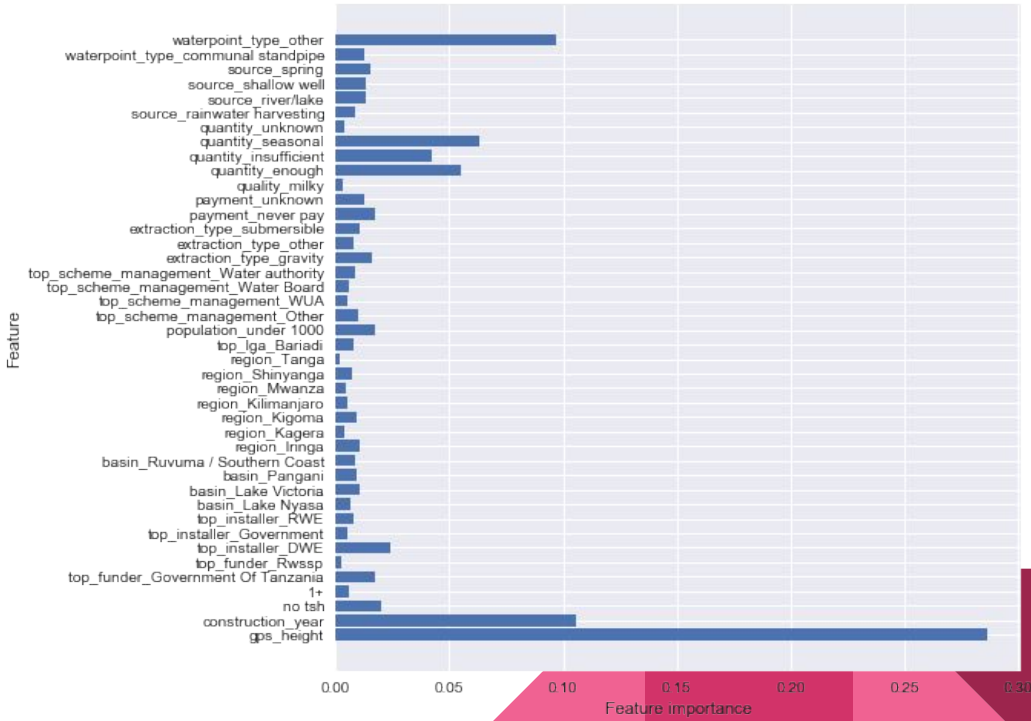
Building a Classification Model

- Boruta algorithm for feature selection
- Used best features in 3 models:
  - Logistic regression: 72.4
  - Decision tree: **75.5**
  - Random forest: 70



Best Features in Model

- GPS height - Water level of the well
- Year the well was constructed
- Type of well



Next Steps

- Prioritize and anticipate maintenance on older wells and uncommon types of wells (not communal standpipes or hand pumps)
- Analyze data through a GIS software
  - Perform a Geographically Weighted Regression model

Thank you!