Project 3 - Tanzanian Waterpoint Classification

—

By Htoo Aung Latt

The Project - Goal

- Predicting the conditions of water points to allow repairs

- Insights into the reason why some water points are more easily broken

The Process - OSEMN

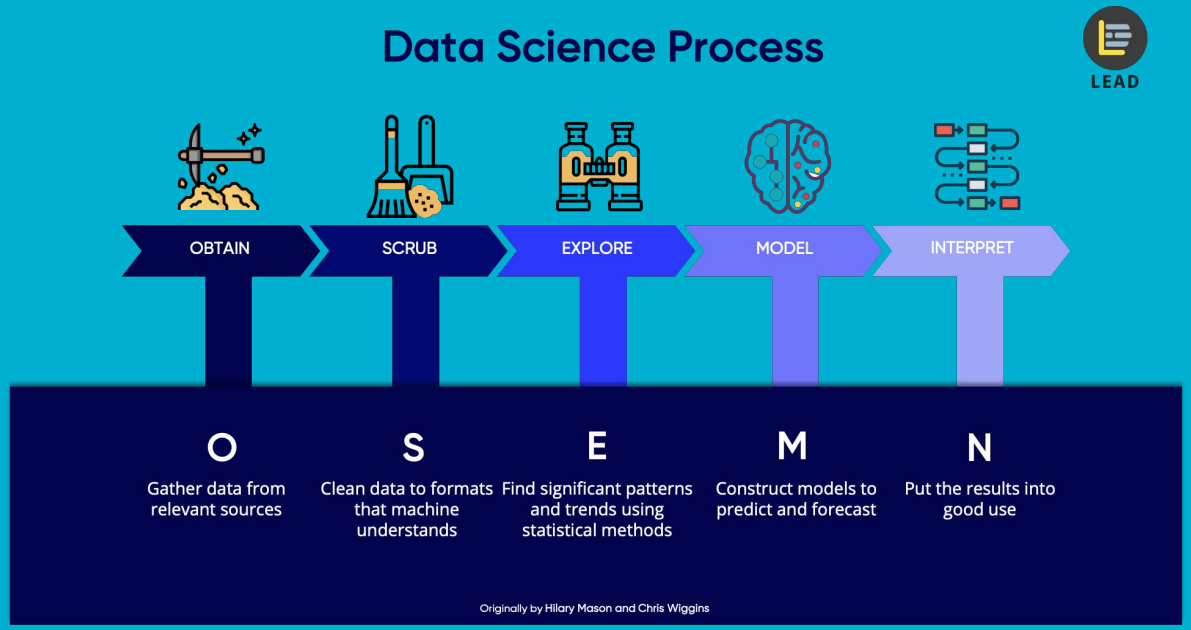
O - **O**btaining the Data

S - **S**crubbing the Data

E - **E**xploring the Data

M - **M**odeling the Data

N - i**N**terpreting the Data



Scrubbing Data

- Columns with identical data.
- Filling missing data with KNN-Imputer.

Filler values found:

- population - 0 ,1
- construction_year - 0
- management_group - unknown
- payment_type - unknown
- water_quality - unknown
- quantity - unknown

- scheme_management - Who operates the waterpoint
- scheme_name - Who operates the waterpoint
- management - How the waterpoint is managed
- management_group - How the waterpoint is managed

- extraction_type - The kind of extraction the waterpoint uses
- extraction_type_group - The kind of extraction the waterpoint uses
- extraction_type_class - The kind of extraction the waterpoint uses

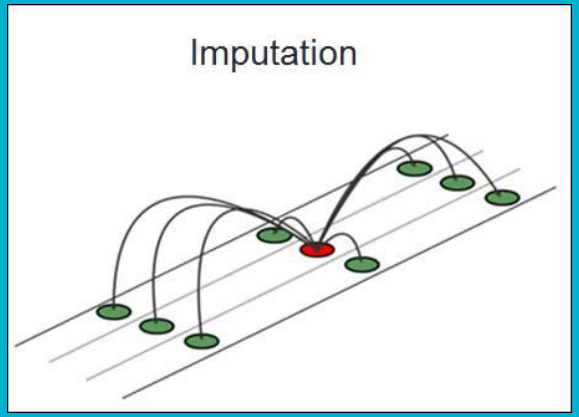
- source - The source of the water
- source_type - The source of the water
- source_class - The source of the water

- water_quality - The quality of the water
- quality_group - The quality of the water

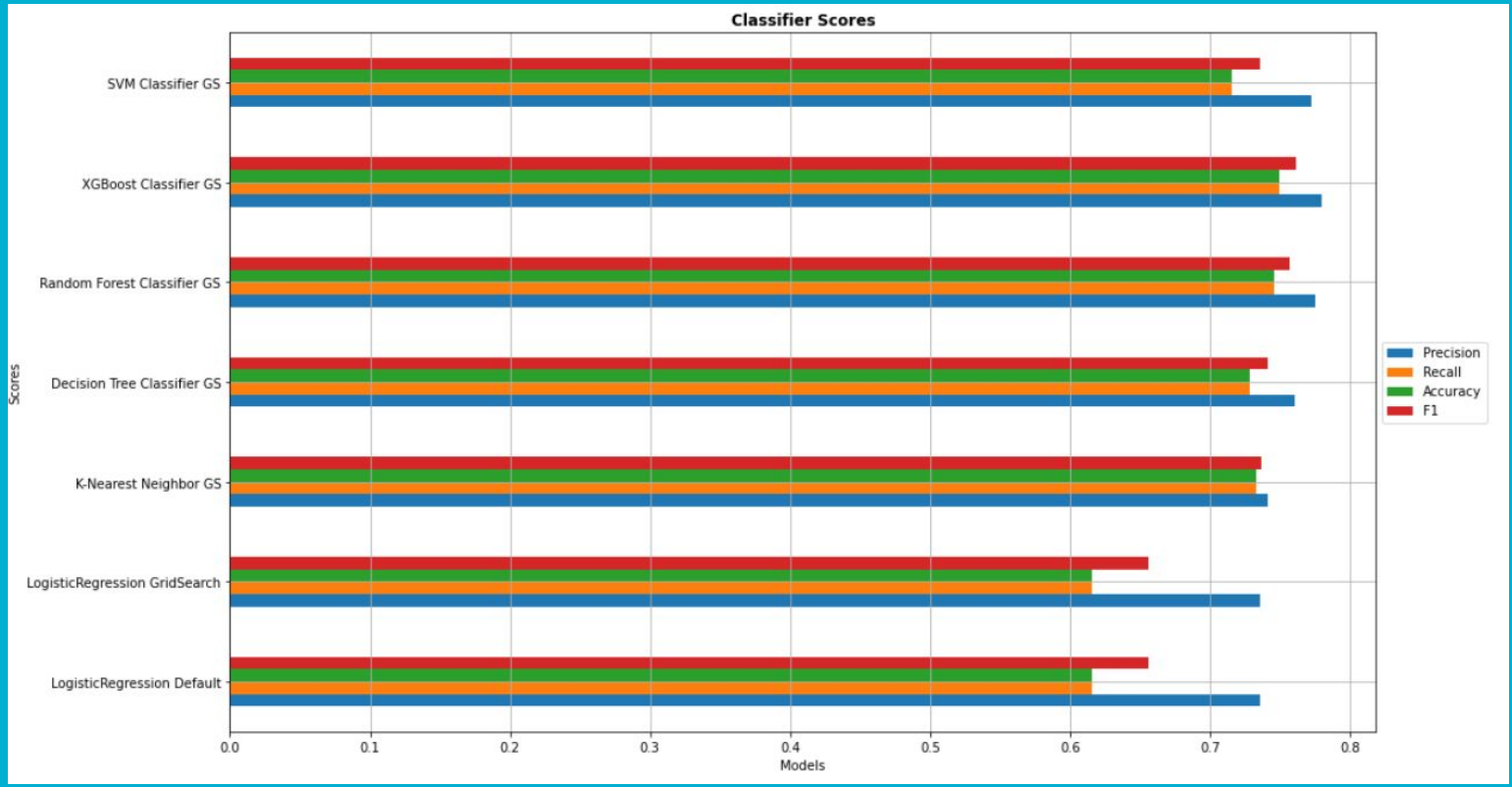
- quantity - The quantity of water
- quantity_group - The quantity of water.

- payment - What the water costs
- payment_type - What the water costs

- waterpoint_type - The kind of waterpoint
- waterpoint_type_group - The kind of waterpoint



Selecting The Model



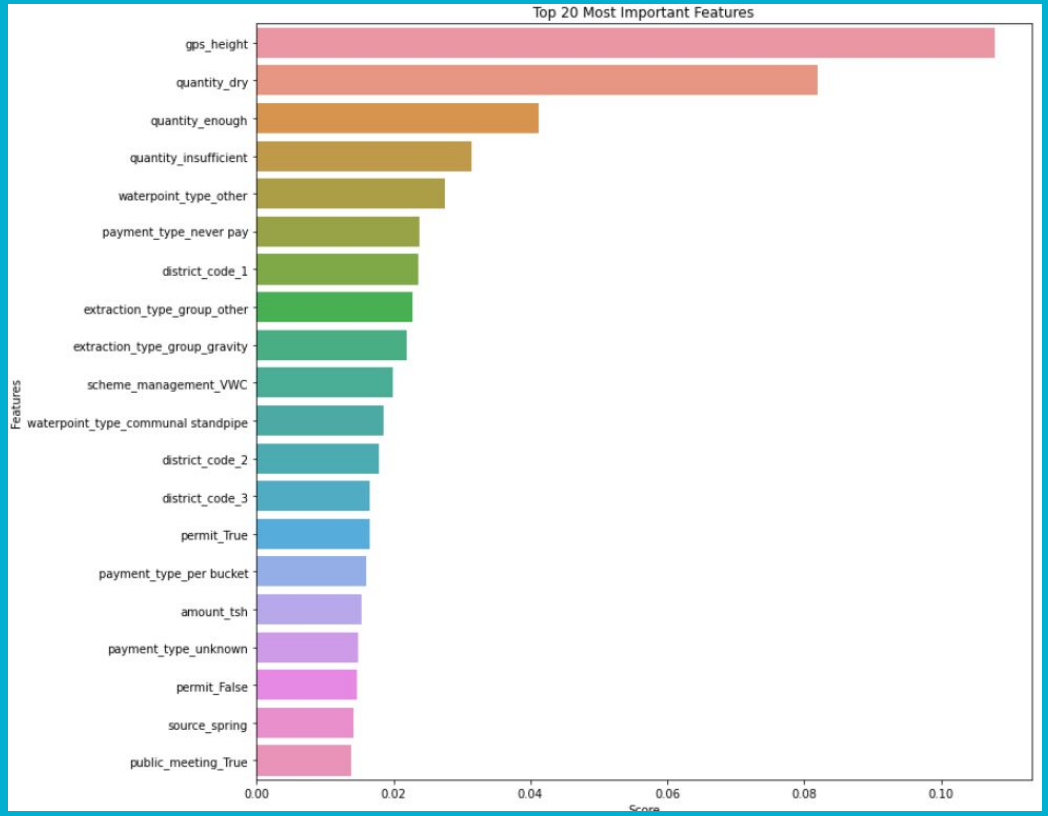
Random Forest vs. XGBoost vs. SVM

Balance between Recall and Accuracy

High recall leads to detecting more functional needs repair category but also raise the false positive. Which would lead to wasted manpower and budget.

Interpreting the Model

- Altitude of well
- Quantity
- Waterpoint type
- Extraction Type
- Payment Type
- District Code
- Public Meeting



Thank You