Pump it Up : Data Mining the Water Table

Deepali Sharma

January, 2023

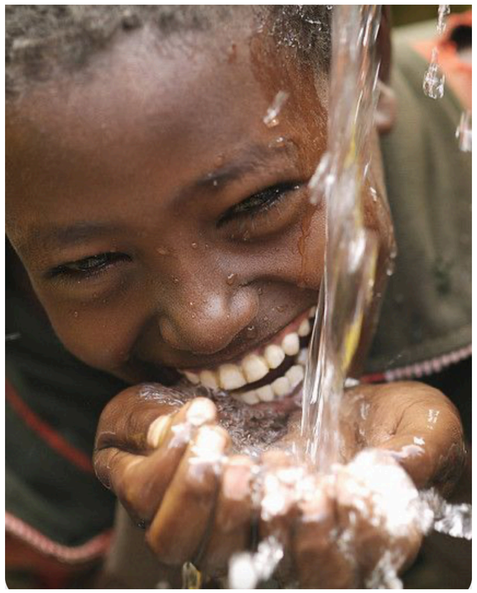
https://www.africanews.com/2022/10/28/tanzania-begins-water-rationing-due-to-drought//

- **Stakeholder:**
  - Ministry of water; Government of Tanzania
- **Business Problem:**
  - Predict the pumps that are functional/faulty/need repairs
  - Proper Allocation of resources based on pump status
  - Find other factors that are responsible for non-functioning pumps

Data:

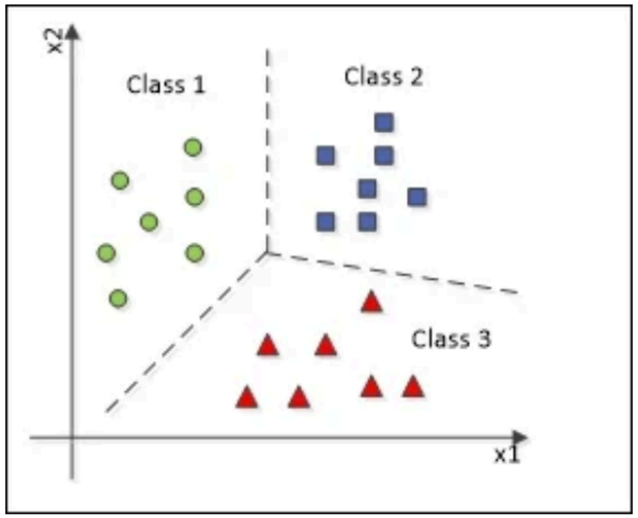
- Data collected by **Taarifa** and **Tanzanian Ministery of Water**
  - Information contains geographical **locations**, **quality** of water, water **extraction** method, installed/funded by etc.etc

Goal:



- **Goal** is to build a **classifier** that would predict the status of water pump utilizing the available information
  - appropriately **redistribute crews** and funds to fix non-functioning pumps
  - where **not to build pumps**, **which extraction techniques** to use etc.

Modeling Algorithm



Classification Problem

- Clean and pre-process the data
- Find the model that best describes the data
- Check the model against validation and test data
- Find the most important features that affect pump functionality
- Use the knowledge gained to make recommendations that will improve the pump status

Results from the Best Model:

- ROC_AUC:
  - **overall** performance between TP and FP rates
- ACCURACY:
  - **correct** predictions
- RECALL:
  - **correct true predictions** for a given class (TP+FN)
- PRECISION:
  - correct true predictions out of all predictions (TP+FP)

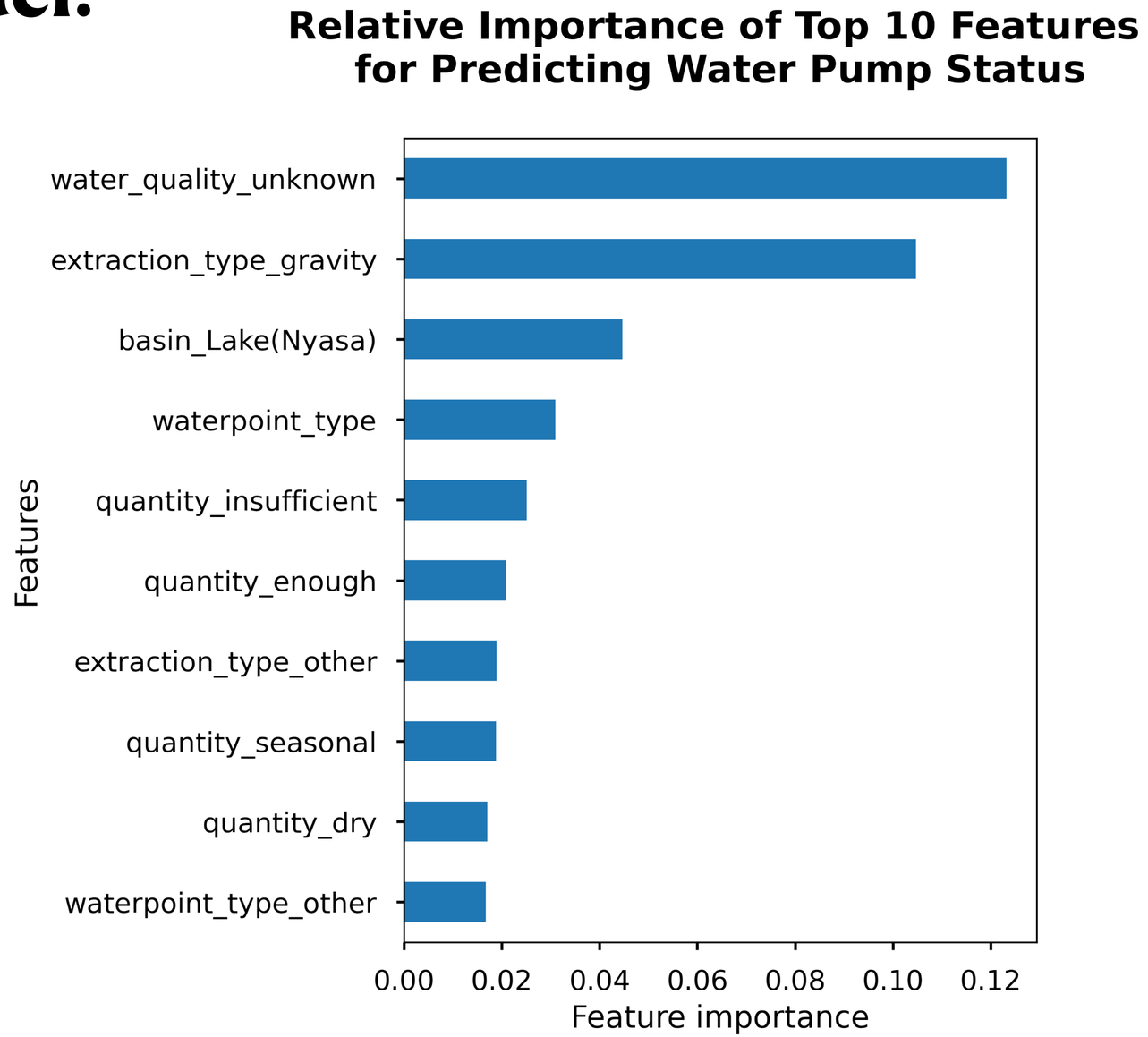
Model Performance:

	TRAIN	TEST
ROC_AUC	91%	88%
ACCURACY	80%	78%
RECALL	78%	78%
PRECISION	80%	78%

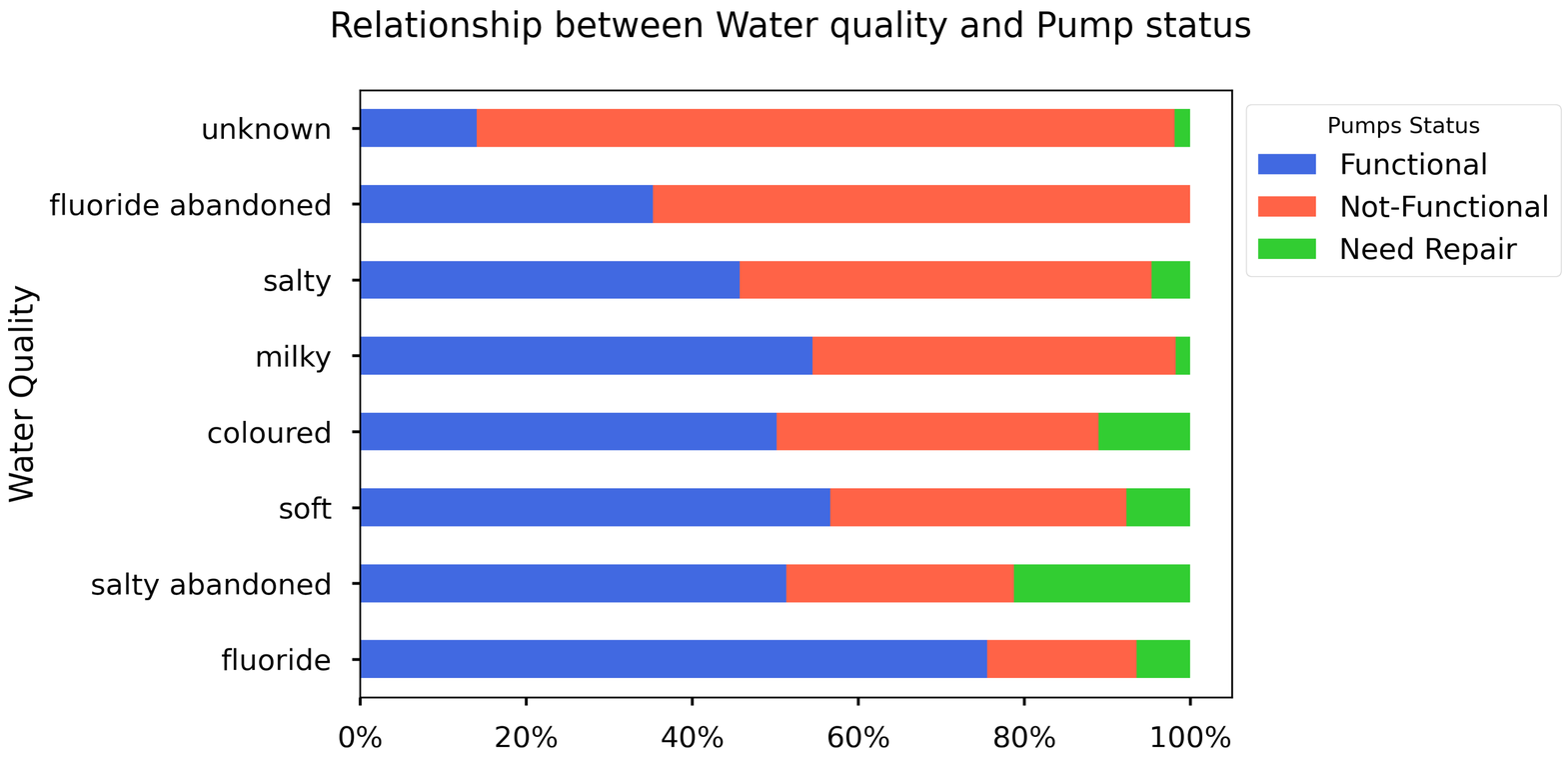
**FP(False Positives), TP(True Positives), FNFalse Negatives), TN(True Negatives)*

Results from the Best Model:

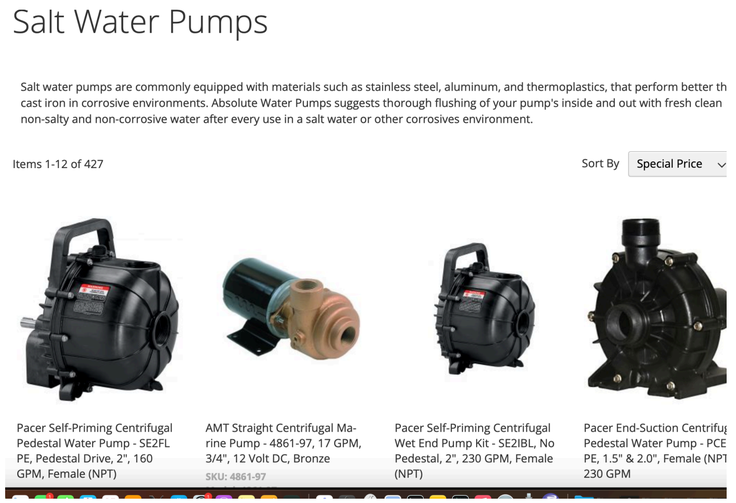
- **Top 3 Features**
  - Water Quality
  - Extraction Type
  - WaterPoint Type



Water Quality: the top feature

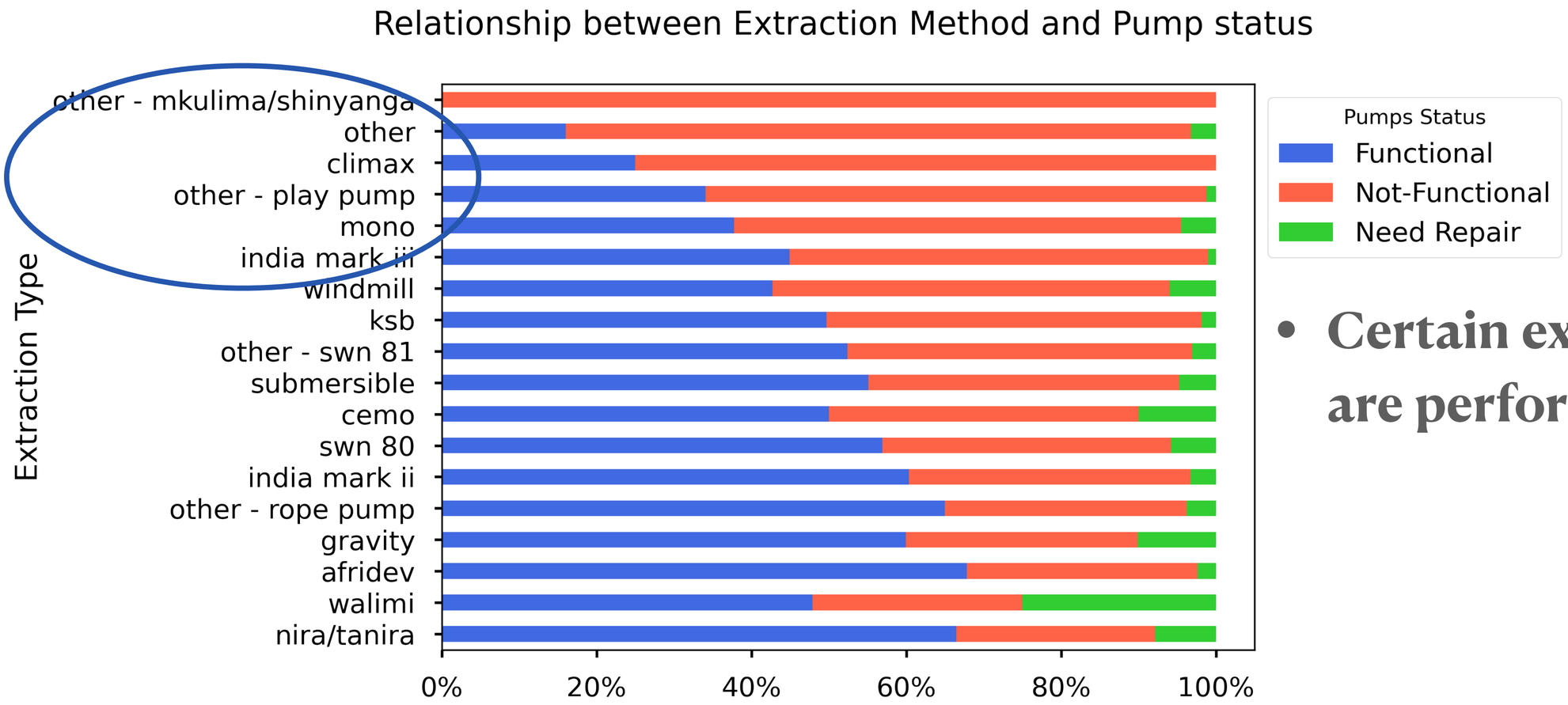


- **salty** water and **unknown** type water pumps fall in the category of non-functional/ need repairs category



- **prioritize** fixing water pumps with **appropriate type** e.g. aluminum pumps for salty locations!

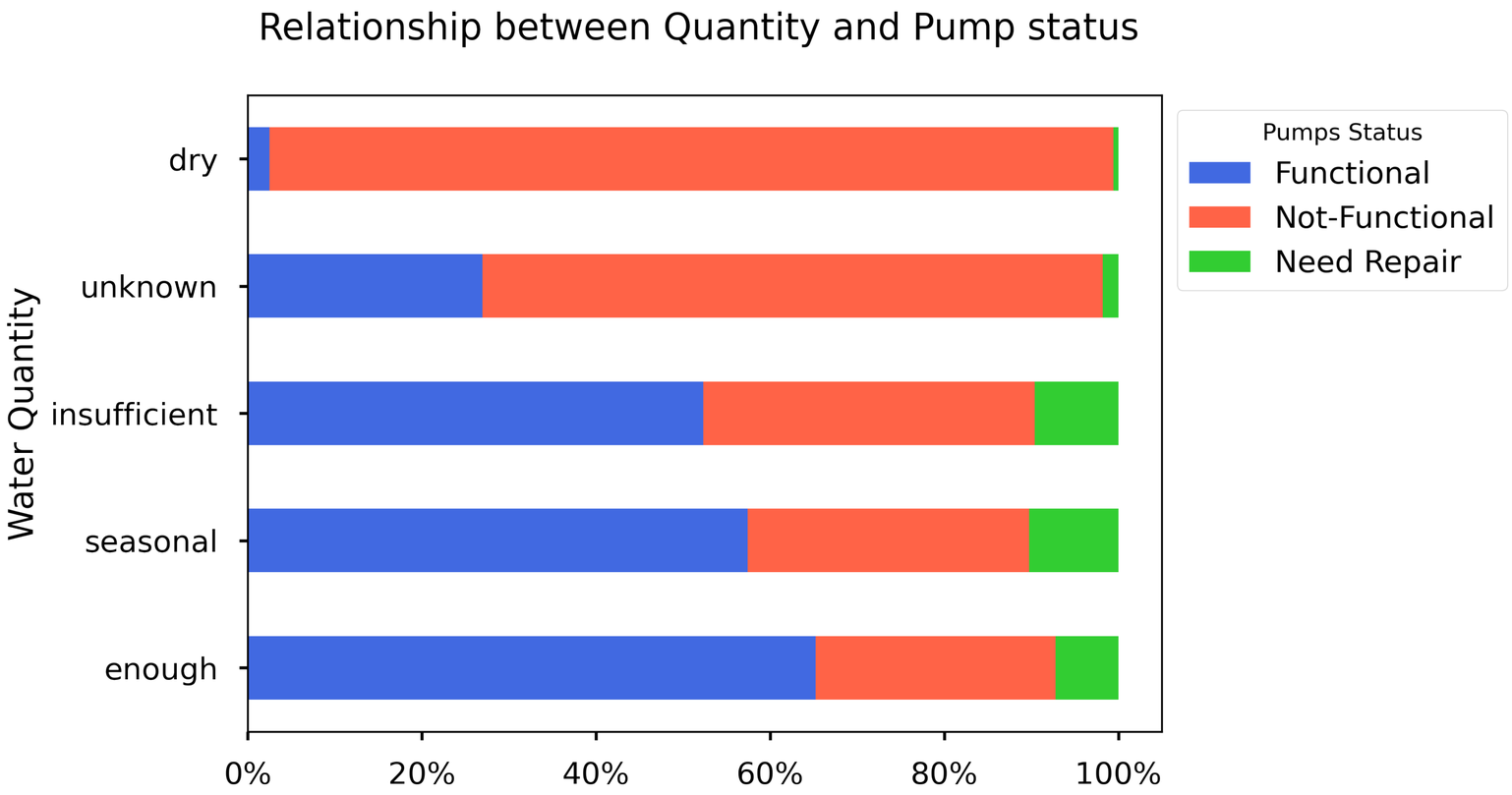
Extraction type: other important feature



- **Certain extraction types are performing worse!**

- Only invest in the extraction techniques that are robust!

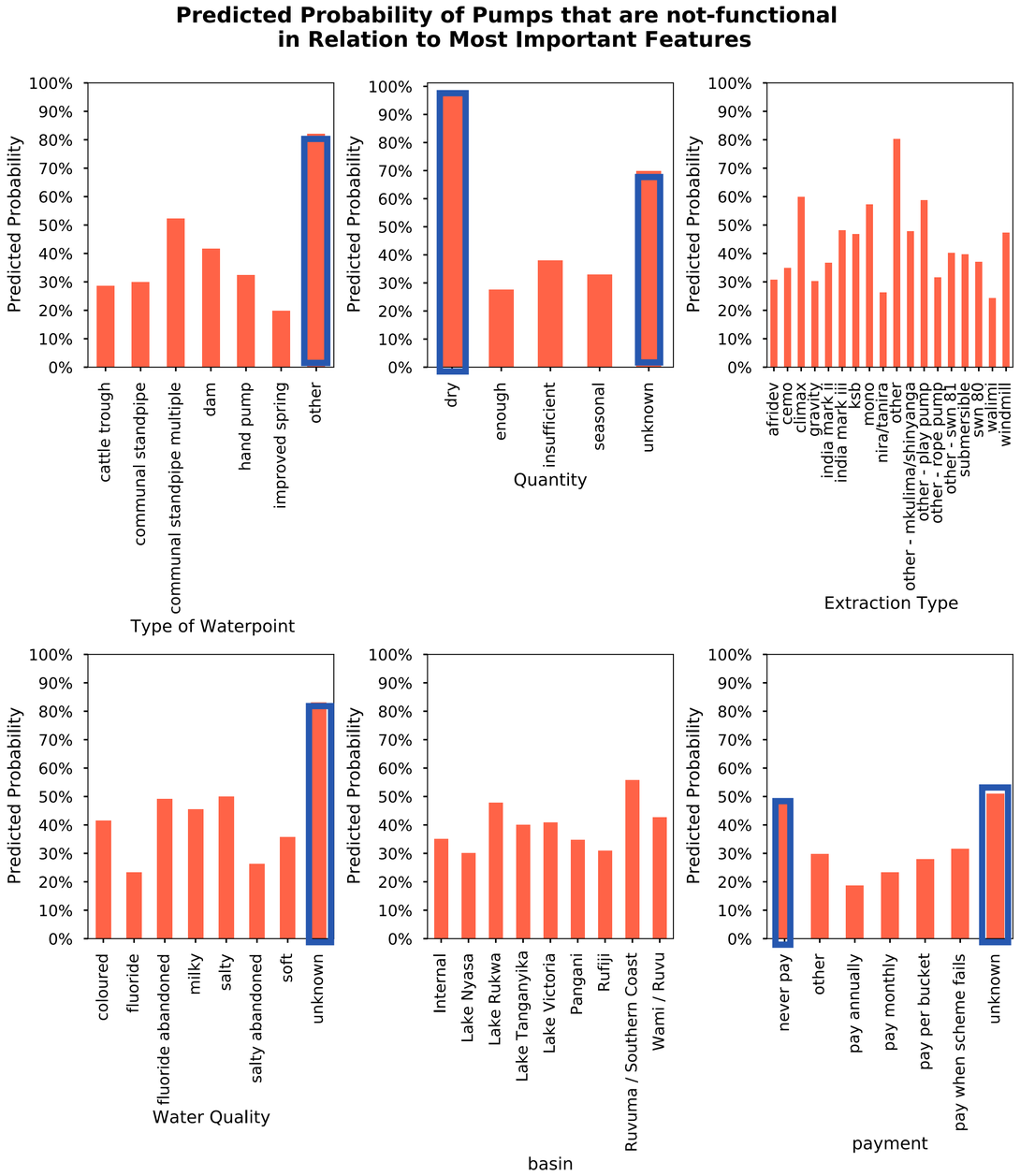
Water quantity: another important feature



- Pumps in **dry region** are **quoted** as **non-functional** or just labeled as such.
- Large fraction of **unknown pumps** are also non-functional

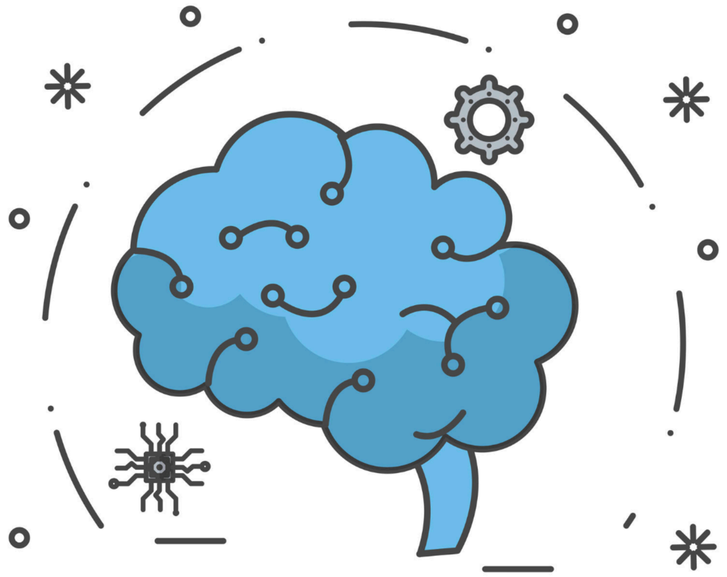
- **Investigate the pumps in dry and unknown categories!**

Predicted Probabilities



- **Predicted probabilities for non-functional pumps are higher among these categories**
  - Pumps types:
    - **other**
  - Quantity:
    - **dry** or **unknown** quantity
  - Water quality:
    - **unknown** followed by **salty**
  - Payment:
    - **never pay** or **unknown**

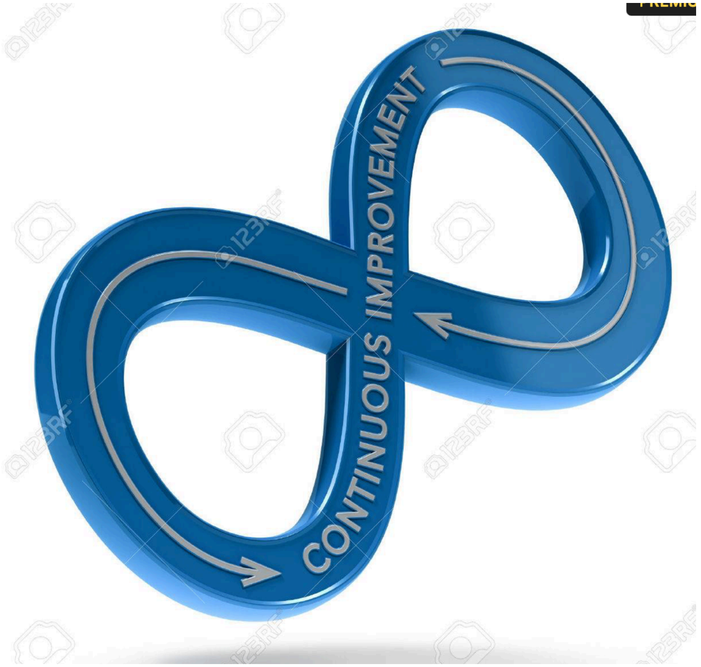
Some Recommendations:



Recommendations

- **Reallocate funds** to replace pumps at **salty** locations.
- **Budgets for R&D** to find out the right extraction type for a given location.
- Why for some pumps **payments are never paid** or **unknown.**
- Look into the water pumps that are in **dry areas** and or where the **quantity** of water is **not known**

Limitations and Improvements



- **Clever strategy** to clean up some of the columns
- **Optimize GridSearchCV** running time
- Check other **classification models like CNN** to see their performance.
- Explore the **dependence on the scoring metric** used in the GridSearchCV
- **Eliminate or quantify effect of parameters** such as installer, funder, scheme_name on the model performance

- **Deepali Sharma**: email:(deeps.sharma@gmail.com, deepali@rcf.rhic.bnl.gov)
- **Linkedin:** https://www.linkedin.com/in/deepali-sharma-a83a126/
- **GitHub**: https://github.com/deepssharma