Pump it Up : Data Mining the Water Table

Deepali Sharma

January, 2023

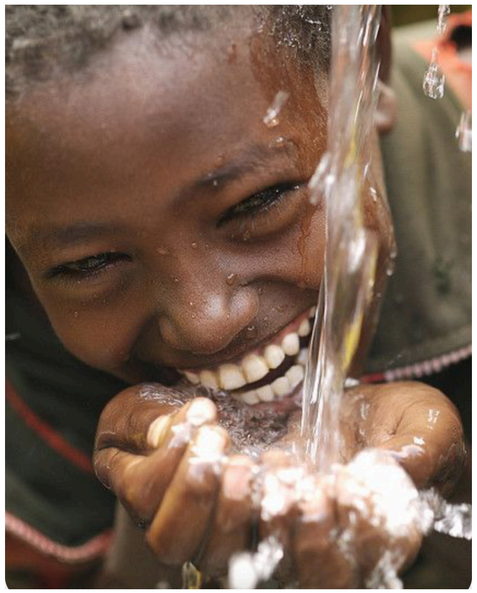
Tanzania begins water rationing due to drought

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

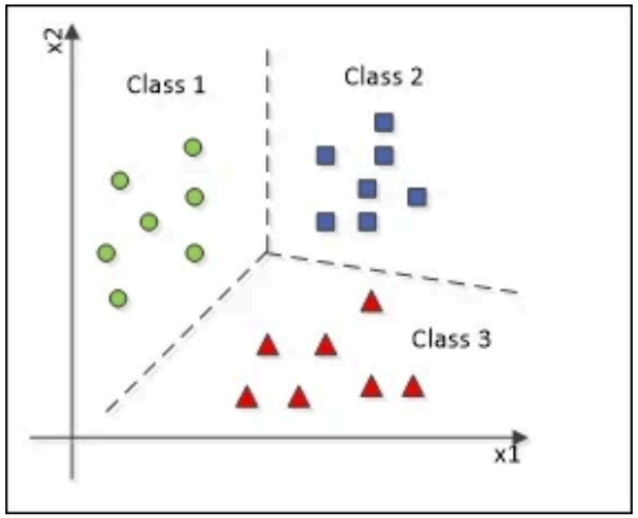
Data:

Goal:



- Data collected by **Taarifa** and **Tanzanian Ministery of Water**
  - Information is provided for water pumps that contain geographical locations, quality of water, water extraction method, installed/funded by etc.etc
- Goal is to build a **classifier** that would predict the status of water pump utilizing the available information
  - Using the predicted information about water pumps, Tanzanian Govt. can appropriately redistribute their crews and funds to fix non-functioning pumps
  - Find the most important features affecting the pump quality that provide insights into e.g, where not to build pumps, which extraction techniques to use etc.
  - Ultimate goal is to provide clean water access to everyone

Modeling Algorithm



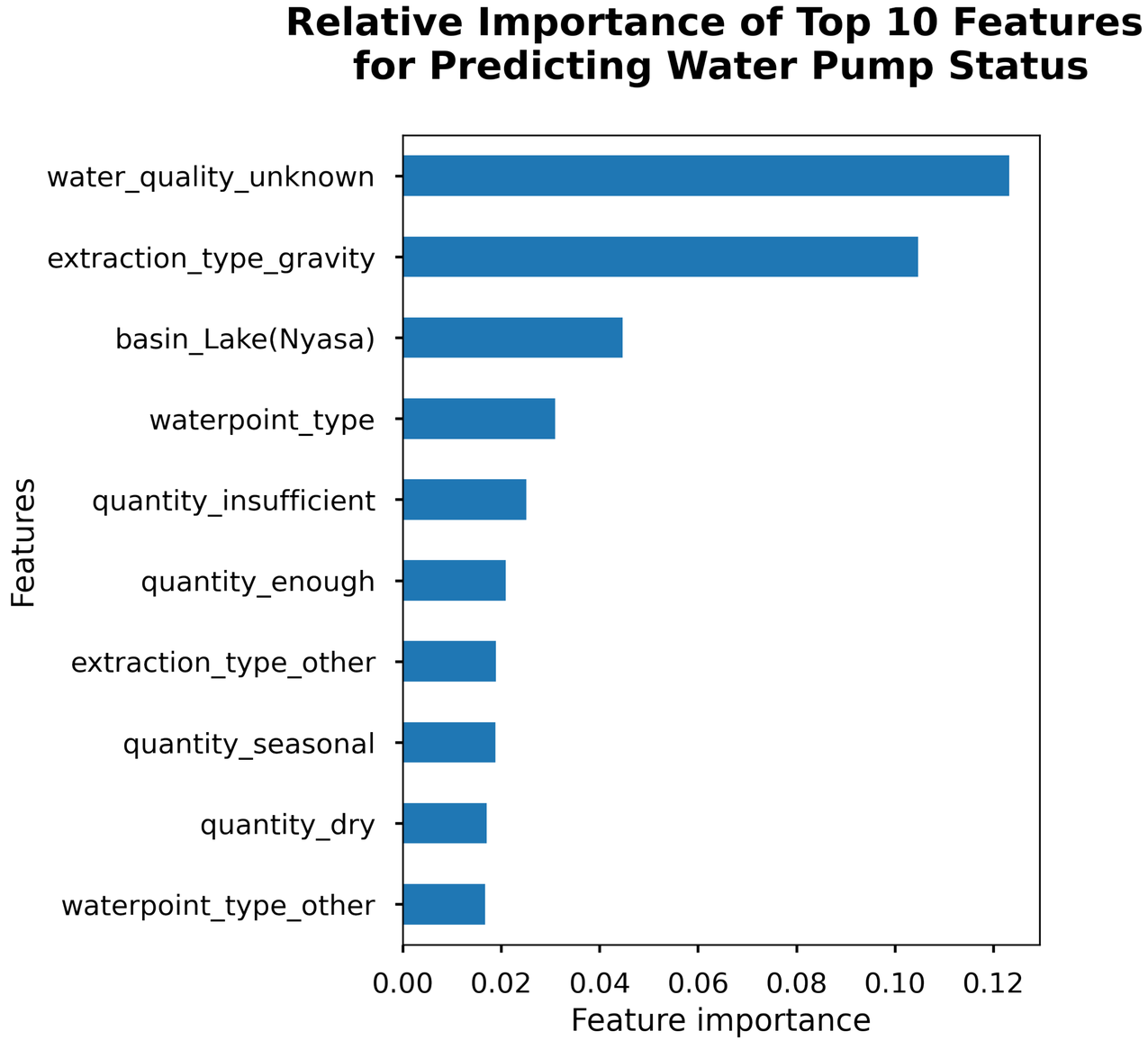
Classification Problem

- 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

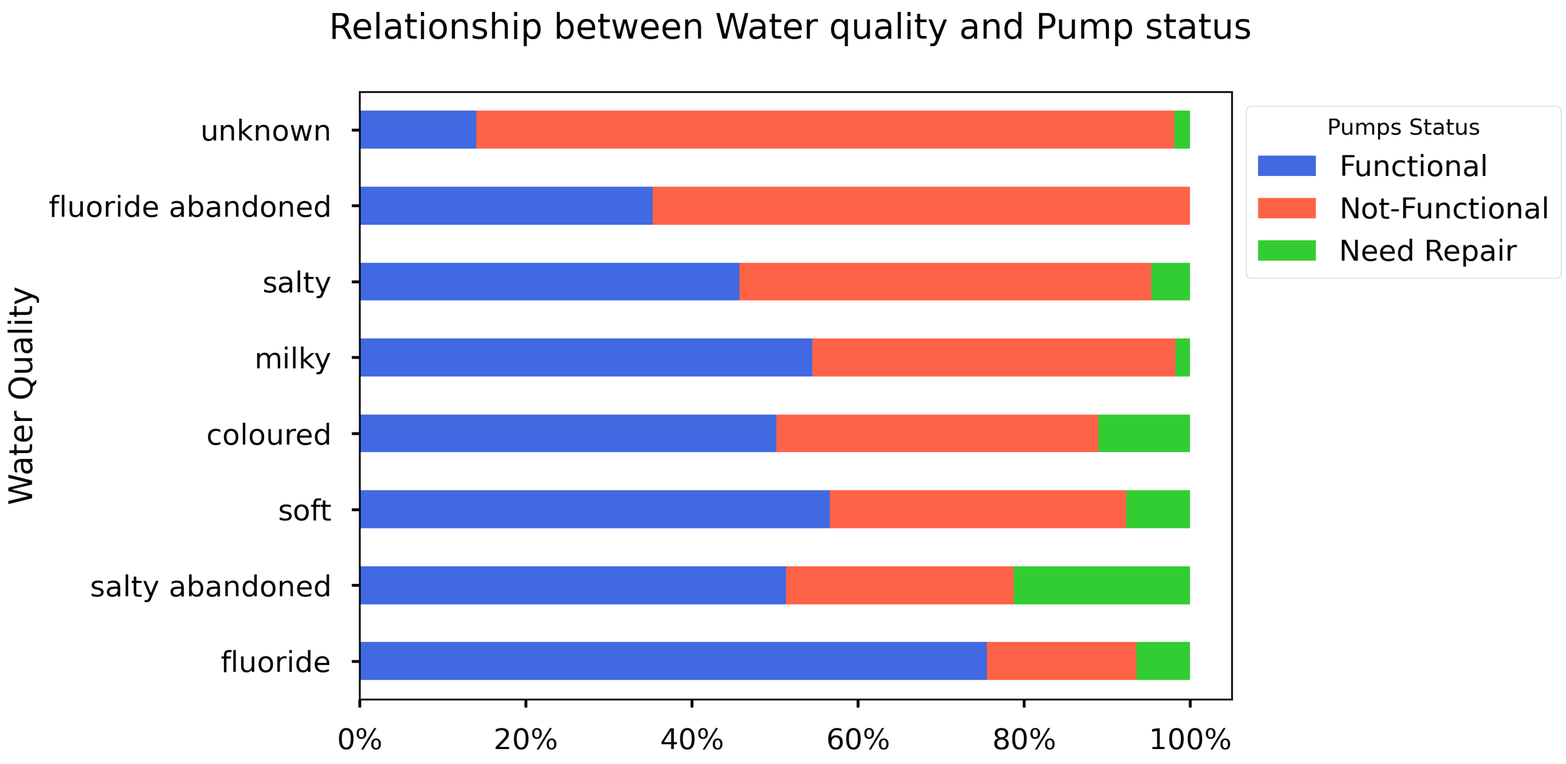
Results from the Best Model:

Model Performance:

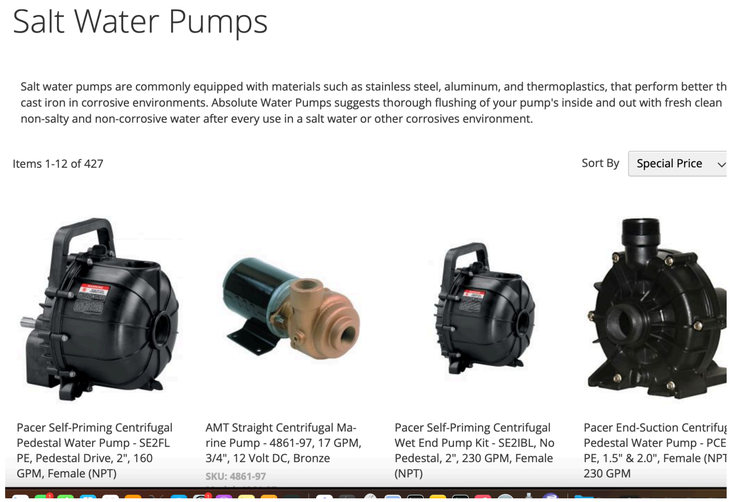
	TRAIN	TEST
ROC_AUC	91%	88%
ACCURACY	80%	78%
RECALL	78%	78%
PRECISION	80%	78%



Water Quality: the top feature

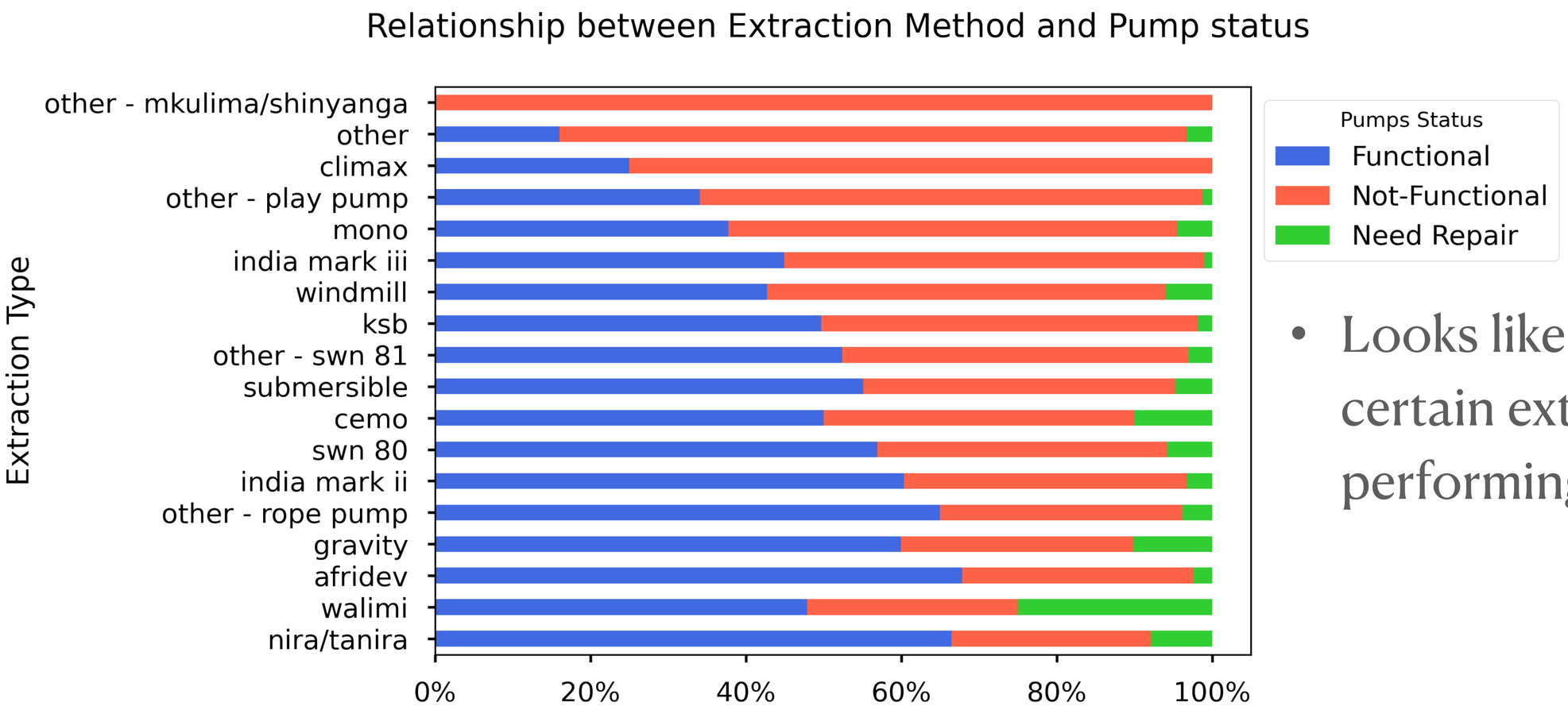


- Most of the **salty** water and **unknown** type water pumps fall in the category of non-functional/need repairs category



- The government should prioritize fixing those water pumps with appropriate types such as Aluminum pumps for salty locations!

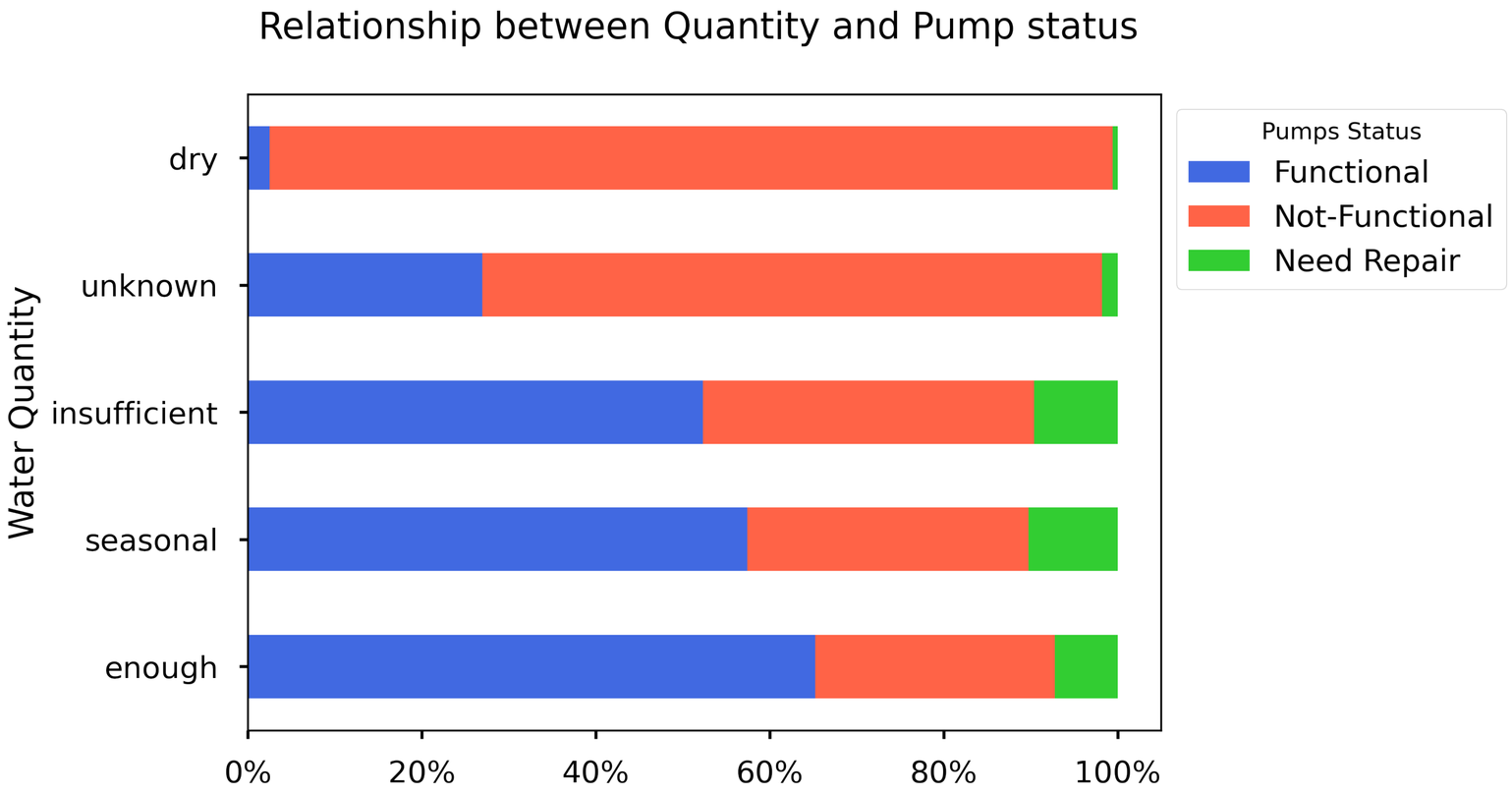
Extraction type: other important feature



- Looks like the pumps with certain extraction types are performing worse!

- The government should only invest in the extraction techniques that are robust!

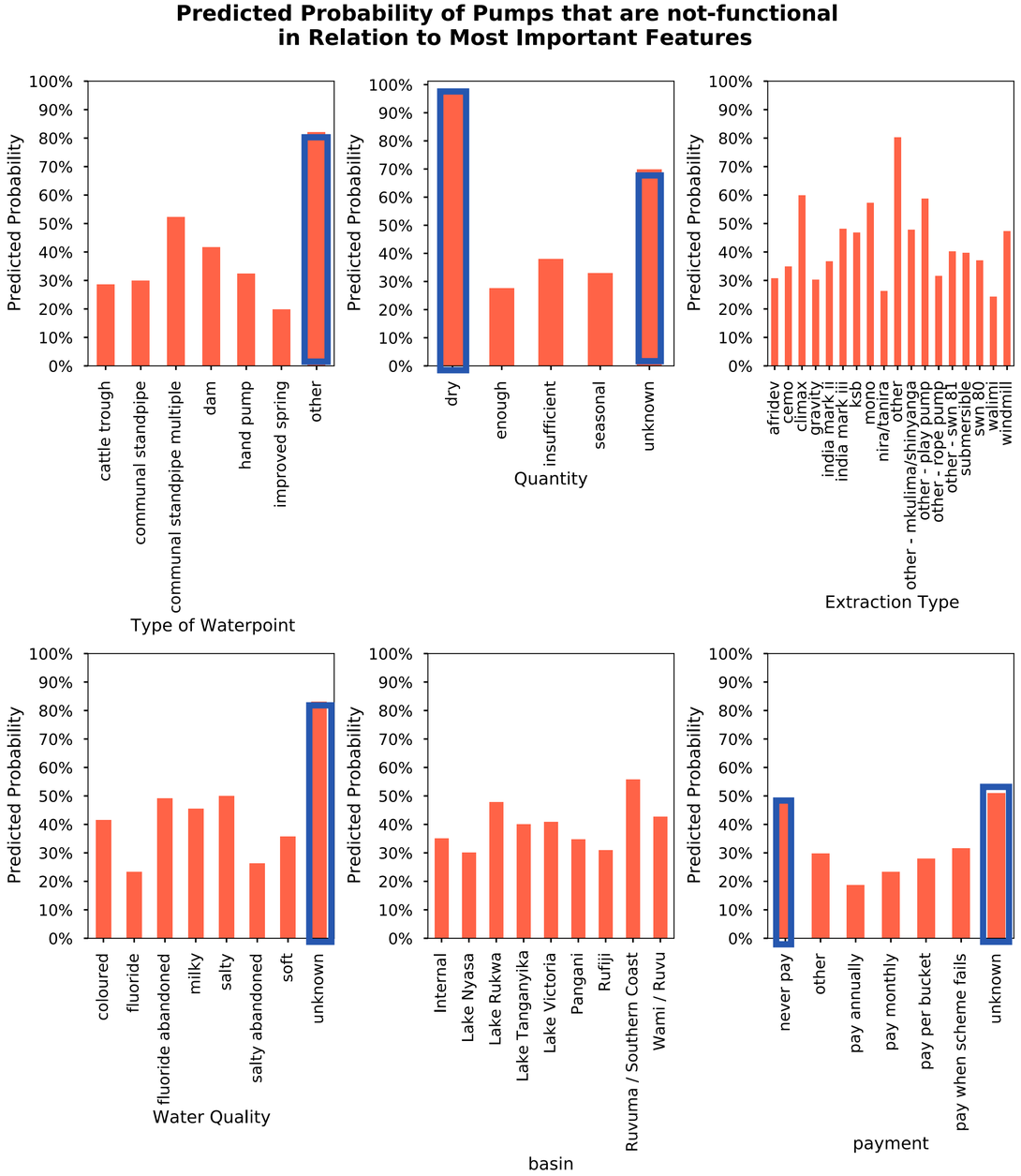
Water quantity: another important feature



- Clearly the pumps in dry region are quoted as non-functional or just labeled as such since no water
- A large fraction of unknown pumps are also non-functional

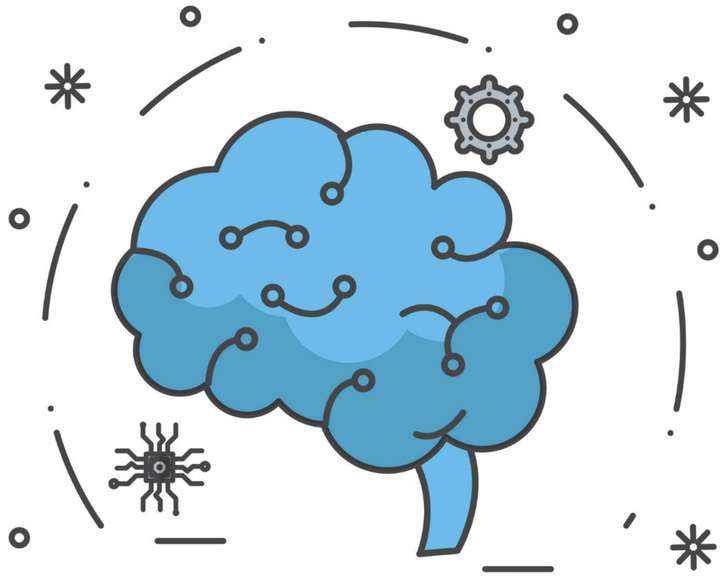
- Investigate the pumps in dry and unknown categories!

Predicted Probabilities



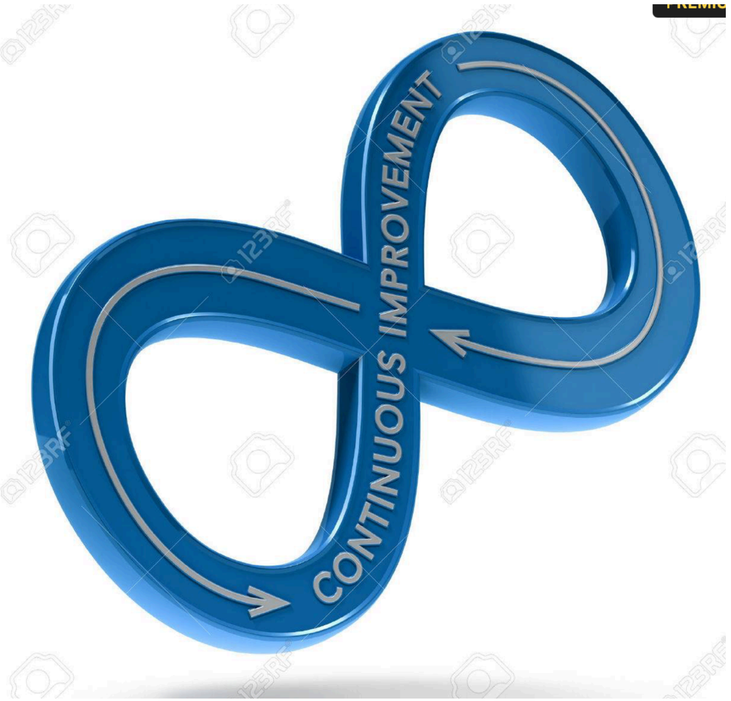
- The predicted probabilities for non-functional pumps are higher among these categories
  - Water pumps types labeled as **other**
  - Water pumps which are **dry** or have **unknown** quantity
  - Water quality **unknown** followed by **salty**
  - And payment is specified as **never pay** or unknown

Some Recommendations:



- The Water ministry should reallocate funds to replacing pumps at salty locations with the correct types.
- Allocate some budgets for R&D to find out the right extraction type for a given location.
- Investigate for the reasons why for some pumps payments are never paid or unknown.
- Also look into the water pumps that are in dry areas and or whom the quantity of water is not known

Limitations and Improvements



- Come up with better strategy to clean up some of the columns (installer, funder and some more)
- Find out if I can run GridSearchCV in reasonable amount of time
- Try some more classification models like CNN to see if they perform any better
- Explore if there is any dependence on the scoring metric used in theGridSearchCV
- Eliminate or quantify if the parameters such as installer, funder, scheme_name have effect on improving 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