Project Proposal

Problem Identification:

- Problem Statement Formation:

  How can we help predict the failure of water points spread across Tanzania before they occur, using a ML model or models that perform better than a baseline model?
- Context:

  There is a huge number of pumps(i.e.,~60K) that provides clean, potable water to communities spread across Tanzania. Compared to other infrastructure projects, water pump projects consist of a high number of inspection points that are geographically spread out. This means, maintenance of such a project is not only time consuming, but also labor intensive, and hence, expensive. However, water projects are arguably one of the most essential projects. For a stakeholder like the Tanzanian Government, specifically the Ministry of Water, it is imperative that they dedicate as much resources as possible so that it's citizens do not face water shortages, and all other issues such as diseases that follow.

  Since resources are limited, being able to predict which pumps are at the risk of failure ahead of time can help minimize the overall number of pump failures, while maximizing the inspection and maximization of water pumps.
- Criteria for success
  - Develop and measure the performance of classification models, to estimate the probability of water points to be in one of three predefined states of 'Functional','Functional, but needs repairs', and 'Not Functional'.
- Scope of solution space
  - Multiple models will be developed.
  - Explainability/Interpretability Study with feature importances will be conducted.
- Constraints
  - The data-dictionary is not very descriptive, and hence we have to make assumptions about the various features, and the data they contain.
- Stakeholders
  - Tanzanian Ministry of Water
  - Local Governments
  - Scholars
  - Researchers
  - NGOs/ INGOs
- Data sources
  - drivendata.org [Link]
    - Taarifa [Link]
    - Tanzanian Ministry of Water [Link]

Approach to the problem:

- Build models, for which I envision that the following points must be addressed, and probably more:
    - Start with something simpler such as Logistic Regression (with more interpretability). Also, try out Tree based algorithms. Compare results with boosting technique based algorithms, specifically CatBoost, since most of the features in our data set are high cardinality categorical variables.
    - The target variable currently has three classes with a significantly small amount of the third class.
        - The tentative plan is to reduce the tree classes to two
        - This might be done by either completely excluding the observations with the third class
        - or
        - Merging the third class to one of the two bigger(in terms of # of observations)
- If possible, also deliver on other insights.
- The deliverables will include
    - All Jupyter notebooks developed, which will be made available from my GitHub repository,
    - A presentation slidedeck, and
    - A written project report--both also available as PDF documents through my GitHub repository

---
