DSF-PT07: Tanzania Wells

Phase 3 Project
Nduku Kiteng’e

OUTLINE

1. Problem Statement
2. Data Process
3. Model Review
4. Limitations
5. Recommendations

Problem Statement

The goal of this project is to :

1. Predict which pumps are functional, which need some repairs and which dont work at all.
2. Predict one of these three classes based on a number of variables about what kind of pump is operating, when it was installed, and how it is managed.
3. Provide a smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

The challenge from DrivenData. (2015). Pump it Up: Data Mining the Water Table. Retrieved [Month Day Year] at https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table with data from Taarifa and the Tanzanian Ministry of Water.

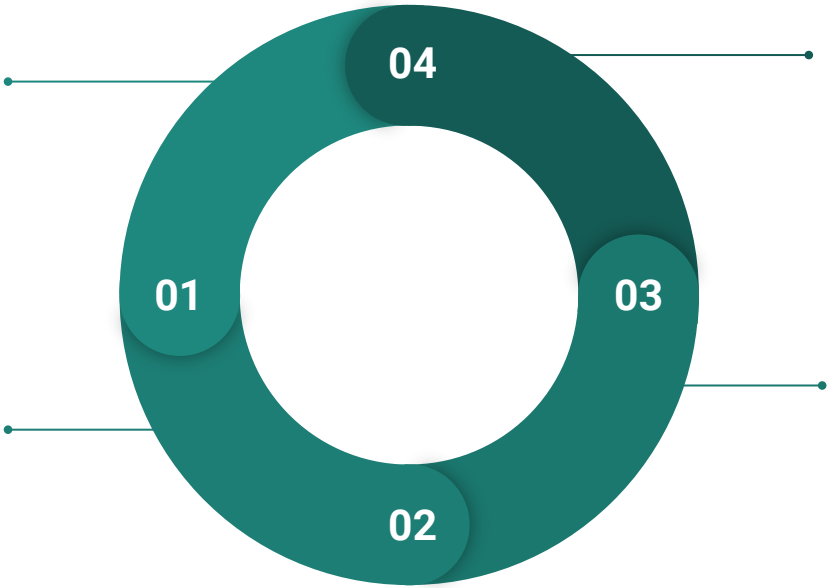
The Data Process

Data Understanding

Session to load and review the data to understand the information it can provide.

Data Preparation

Includes data cleaning for missing values, data types and null values and feature engineering for potential analysis methods and variables.



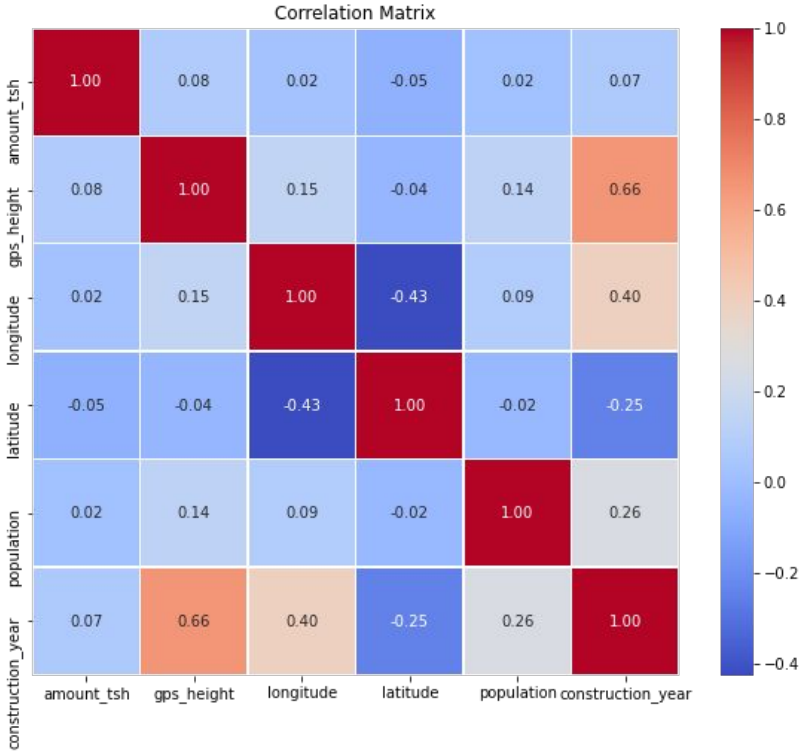
Recommendations

Develop a list of recommendations for the clients especially potential buyers and home owners.

Modelling and interpretation

Using the cleaned and prepared data, we review the potential models based on the identified objectives.

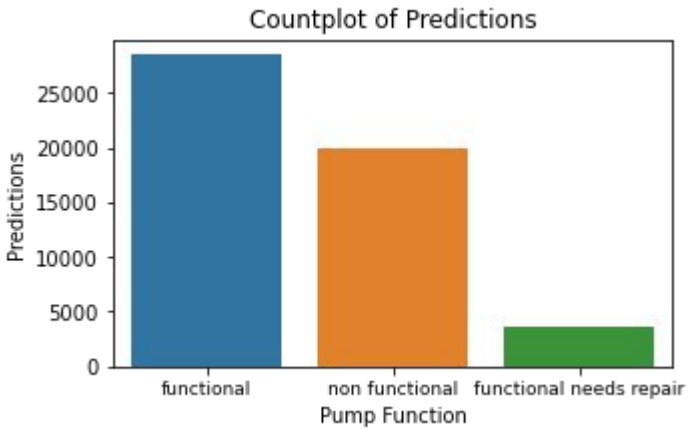
Data Understanding



The following activities were carried out:

- Identify the problem to be solved
- Import necessary libraries (pandas, numpy, sklearn, matplotlib, seaborn)
- Load the dataset
- Set random seed for reproducibility

Data Preparation



The following activities were carried out:

- Examine the dataset (head, info, describe)

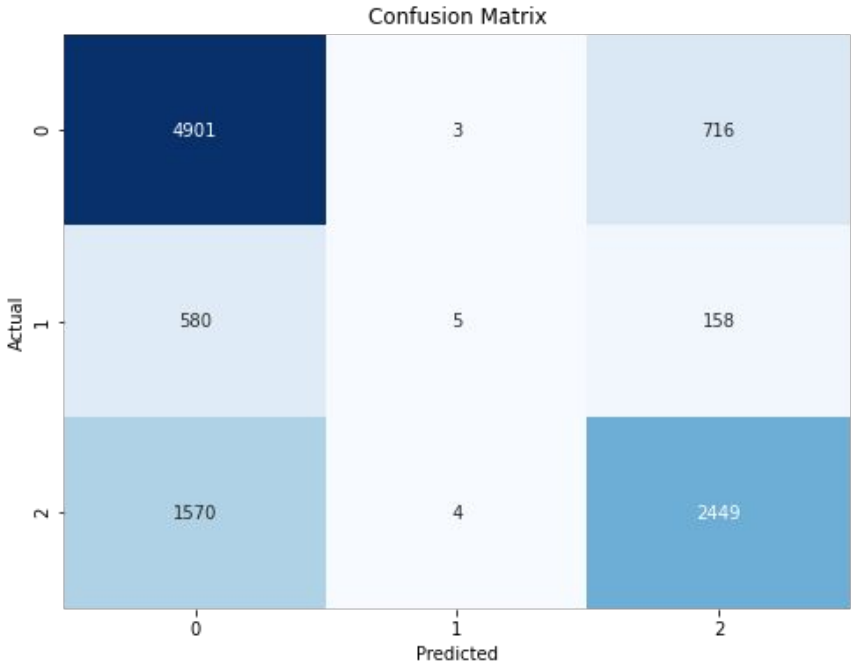
- Check for missing values and handle them

- Explore data distributions and correlations

- Perform feature engineering if necessary

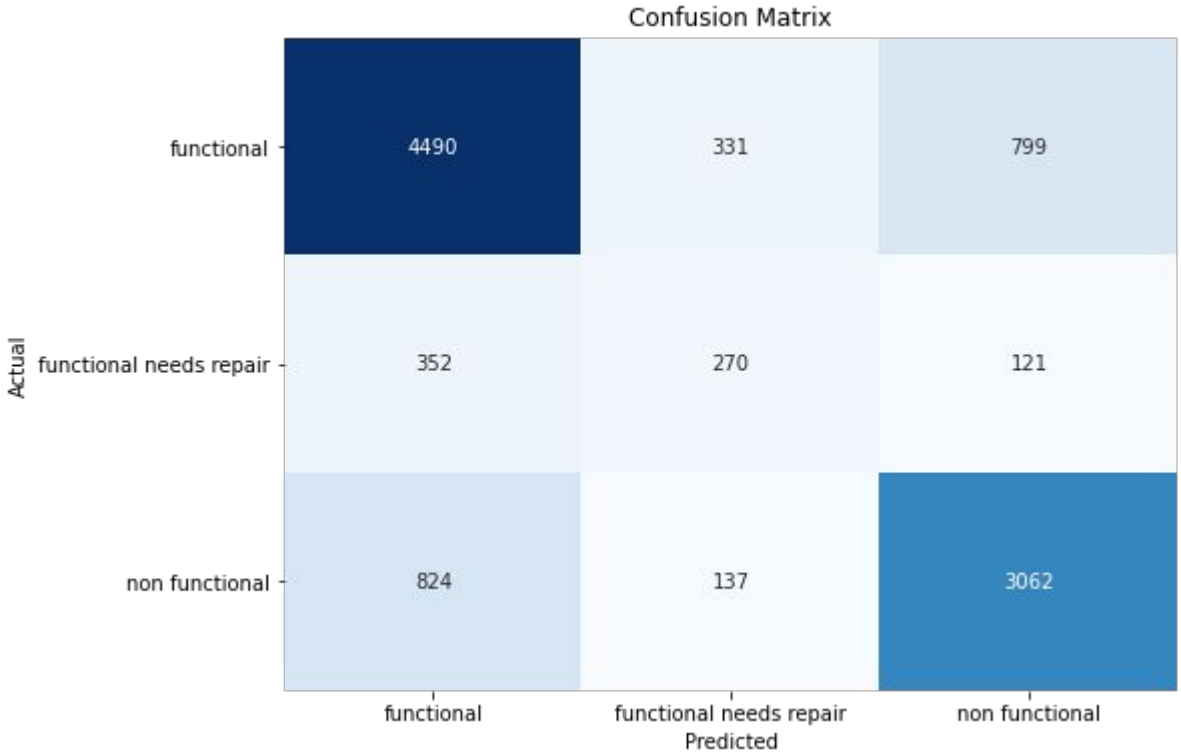
- Split the data into features (X) and target variable(s) (y)

Modelling: Logistic regression



This model correctly classifies about **70.8%** of the cases in the test set, which is moderately good but shows room for improvement.

Modelling: Decision Trees



75.31% of the water pumps were correctly classified as either functional, needing repair, or non-functional. While this is a decent result, it can still improve, particularly in distinguishing between the different classes.

Model Comparison

	Metric	Decision Tree	Logistic Regression
0	Accuracy	0.757	0.708
1	Functional Precision	0.790	0.700
2	Functional Recall	0.800	0.870
3	Functional F1-Score	0.800	0.770
4	Functional Needs Repair Precision	0.370	0.420
5	Functional Needs Repair Recall	0.350	0.010
6	Functional Needs Repair F1-Score	0.360	0.010
7	Non-functional Precision	0.770	0.740
8	Non-functional Recall	0.770	0.610
9	Non-functional F1-Score	0.770	0.670

The Decision Tree (75.8%) outperforms Logistic Regression (70.8%) in terms of accuracy.

- It outperforms Logistic Regression in overall accuracy.

- It handles the "functional needs repair" class significantly better.

- It captures more complex, non-linear relationships, which are likely present in the dataset.

Thank you.
Questions?