PHASE 3 PROJECT
TANZANIAN WATER PUMP ANALYSIS.

Business Understanding.

The current population of the United Republic of Tanzania is 64,152,291 as of Tuesday, February 7, 2023, based on @Worldometers.info elaboration of the latest United Nation data. 4 million People however lack access to safe water. Access to clean water is a fundemental human necessity and not having that access is a serious health risk to many.

Unicef.org states that as part of its Vision 2025, the Government of Tanzania has pledged to improve sanitation to 95% by 2025. The Tanzanian Gorverment hope to achive this by teaming up with NGOs around the world.

In this project our goal is to create a model that would most accurately predict which water pumps are functional,which need repairs and which need to be completely replaced.

BUSINESS PROBLEM

Using data gathered from Taarifa and the Tanzanian Ministry of Water, we are tasked with analyzing the different features corresponding to functional and non functional water pumps with the goal of creating a model that can predict if a pump needs to be replaced.Through this analysis implementation of actionable plans for fixing and replacing water pumps through out Tanzania will be made.

DATA UNDERSTANDING.

This project's datasets come from Taarifa:an open source platform for the crowd sourced reporting and triaging of infrastructure related issues. The datasets were then downloaded from DriveData.

These datasets contain information about 59,400 water pumps throughout Tanzania.The first dataset contains ID numbers and feature information about each water pump. The second dataset contains ID numbers and pump conditions for each water pump.

Pump Features:

- amount_tsh - Total static head (amount water available to waterpoint)
- date_recorded - The date the row was entered
- funder - Who funded the well
- gps_height - Altitude of the well
- installer - Organization that installed the well
- longitude - GPS coordinate
- latitude - GPS coordinate
- wpt_name - Name of the waterpoint if there is one
- num_private -
- basin - Geographic water basin
- subvillage - Geographic location
- region - Geographic location
- region_code - Geographic location (coded)
- district_code - Geographic location (coded)
- lga - Geographic location
- ward - Geographic location
- population - Population around the well
- public_meeting - True/False
- recorded_by - Group entering this row of data

- scheme_management - Who operates the waterpoint
- scheme_name - Who operates the waterpoint
- permit - If the waterpoint is permitted
- construction_year - Year the waterpoint was constructed
- 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
- management - How the waterpoint is managed
- management_group - How the waterpoint is managed
- payment - What the water costs
- payment_type - What the water costs
- 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
- source - The source of the water
- source_type - The source of the water
- source_class - The source of the water
- waterpoint_type - The kind of waterpoint
- waterpoint_type_group - The kind of waterpoint

Pump Conditions:

- functional - the waterpoint is operational and there are no repairs needed
- functional needs repair - the waterpoint is operational, but needs repairs
- non functional - the waterpoint is not operational

DATA PREPARATION

Before modeling the dat had to be cleaned.The data had missing values which we chose to drop.Then some irrelevant colums had to be dropped as well. After this we noticed the longitude and latitude columns need to be worked on as they were very inconsistent with the true position of Tanzania on the map. The data was free from duplicates. A new age column was created from the 'construction_year' and 'date_recorded' columns.

EXPLORATORY DATA ANALYSIS.

A visualization was created to help us see how the variables affect one another. A heatmap map was also created to show the correlation between variables.

DATA MODELLING.

Several baseline models were created and Random Forest Classifier performed better than all the others.The models created were Random Forest Classifier,Decision Tree Classifier,Logistic Regression and so on. Preprocessing and hyperparameter tuning was preformed and new models created. The Randon Forest Model still performed best with an accuracy of 76%

CONCLUSIONS

The final Random Forest Model shows that we can predict the condition of each water pump with 76% accuracy.

We chose this model because of its priority with classifying False Non_Functional over False Functional. This model is most likely not cost effective because it will prioritize classifying a pump as needing to be replaced over being functional.Because of that prioritization though,this model does provide us with the most humanitarian solution and given the data and our project needs,provides the most useful results.

RECOMMENDATIONS

- Given the above conclusions,the priority should be replacing the pumps that needs replacing as this will go along way in ensuring Tanzania reaches its Vision 2025.
- Water points in densly populated areas should be monitored as this are prone to a lot of wear and tear and serve many people.
- More reseach need to be done to areas with pumps that need replacing so as to establish the real cause before replacing.
- Research should also be carried out on the pumps that require repairs.This is for better understanding on which repairs take priority.