Tanzania Water Wells Classification

Problem Overview

The Tanzania Ministry of Water along with Taarifa, a crowd-source platform, have commisioned the development of a predictive model that is supposed to be able to predict with **water wells** are likely to fail. While much of Tanzanias population has access to basic water services, a large 39% of households still lack this basic need. An estimated 10% of preventable deaths in the country can be attributed to inadequate *wash services*. A predictive model can enable quick **predictive maintenance** on water wells and help ensure water security in many of the rural communities that are disporportionately affected by this problem.

Project Objectives

The main objective of this project undertaking is to build a Classification model that can be able to classify water wells in Tanzania as `functional` or those that need repairs `need_repair`

Specific Objectives

1. To conduct exploratory analysis and determine which features to include in our model
2. Determine the cleaning steps to be included in building the model pipeline
3. To build a classifiction model that can predict the status of wells with acceptable accuracy.

Success Metrics

- `Accuracy` : 75%
- `Recall` : 80%

The Data

The data is provided by an organization known as Taarifa in co-operation with the Tanzanian government.
A detailed description can be found here

Column Summary

- 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

Conclusions

- The Baseline **Logistic Regression** performs well on our classification metrics i.e **accuracy** and **recall** Accuracy is a valid metric as the class imbalance is negligible as seen in the data exploration
- The **KNN model** performs similarly to the Logistic Regression. However, due to its exponentially increasing time complexity, this model has a much longer runtime The *runtime* does not justify using this model.
- Both **Decision Trees** and their ensemble counterpart, **Random Forest** performed worse than the baseline model.
- After tuning the Logistic Regression model, as it preforms best, we obtain the best parameters.
- Using a feature selector proved detrimental to the modelling process.

> - `Best Model:` **logistic regression(tuned)**

Releases

No releases published
Create a new release

Packages

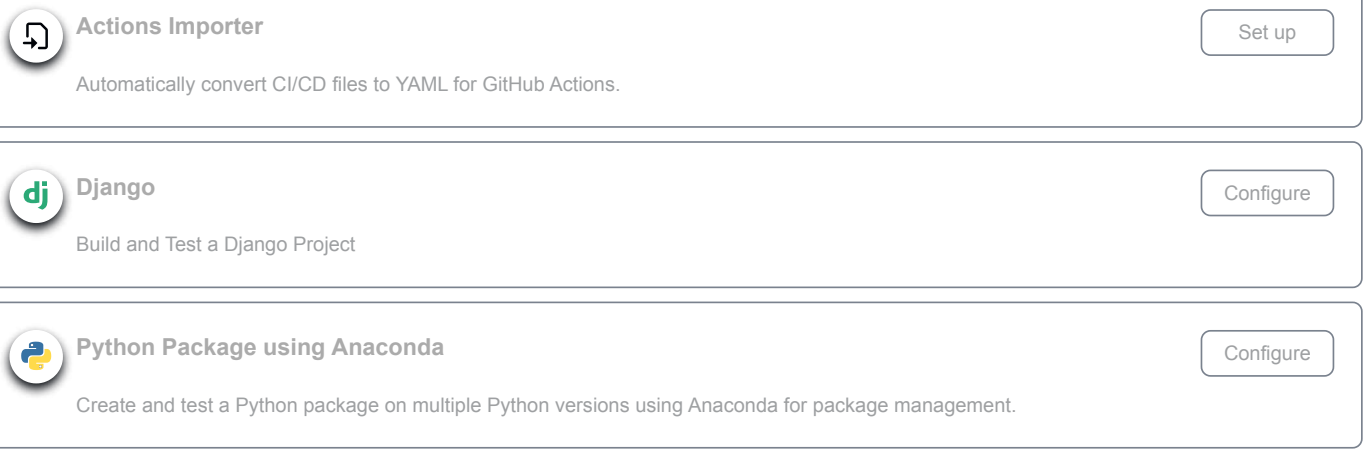
No packages published
Publish your first package

Languages

- **Jupyter Notebook** 99.2%
- **Python** 0.8%

Suggested Workflows

Based on your tech stack



More workflows

Dismiss suggestions