Tanzanian water pump classification

Project Overview

This project

Business Problem

Tanzania is ranked poorly for access to clean water; 70% of the population is rural, where access to water reaches only 45%. The Tanzanian Development Trust is a UK based charitable organization operating within the country of Tanzania since 1975. They focus on development in rural areas, including water accessability and sanitation projects in the northwest regions of Kagera and Kigoma. Repair and rehabilitation of waterpoints is less expensive and more protective of existing water resources, and a new benefactor wants to expand the project to more regions. As a secondary objective, they want to expand the scope of the project to include waterpoint installations where they are needed most: concentrations of failed wells. The task is to build a predictive model to categorize waterpoints as functional, needing repair, or non-functional, as well as map the locations of waterpoints.

The Data

The data is sourced from the Taarifa waterpoint dashboard, aggregated by the Tanzanian Ministry of Water. The data was collected between 2011 and 2013, and contain 59,400 entries. The features include geographic location data, population, source/quality/quantity information, operational data, and more. The data is mostly composed of categorical features; of the 39 raw features, 9 are numerical.

The Process

Initial data exploration revealed null values in 7 features that were addressed in a variety of techniques. Exploration of numerical features showed GPS errors that needed to be fixed as well as some features that would not be modeled and were dropped from the dataframe. Further exploration resulted in engineering three of the numerical features to include in modeling. Categorical features were found to contain some duplicates as well as features not suited to modeling. Other groups of features represented varying degrees of organization for the same thing, like extraction (represented by type, group and class), that we could futher cull from our model. The final dataframe used for modeling contains 18 features.

The modeling data was encoded, train/test split, and scaled. Baseline modeling was performed with a DummyClassifier which resulted in an accuracy of 54%. Three modeling techniques were explored: KNearestNeighbors, RandomForest, and XGBoost. For each of the models, we began with a simple baseline, performed a grid search, and iteratively moved through modeling with optimized hyperparameters as well as techniques for accounty for imbalanced target data. Comparison was between version of models and different models was made using classification reports and confusion matricies.

Conclusions

The XGBoost model using weight scaling was selected for production. While other models performed better when evaluated on total accuracy, the XGBoost model had the highest recall for waterpoints needing repair. Errors made by this model skewed towards classification as 'needs repair, which given the parameters of the program expansion is acceptable. Research during the project showed that well maintenance and repair is imperetive to continued functionality, so mislabeling a functional well as needing repair would simply result some likely necessary upkeep. Exploring the feature importance showed that the quantity of water has the most impact on the functional status of the waterpoint. Other impactful features include the physical description (location and height) and extraction type.

Using the model to predict on the test data, we were able to recommend expansion of well repair operations to the neighboring region of Shinyanga. There is also a concentration of wells needing repair in the southern region of Mbeya if the TDT wanted to continue to expand the program. In addition, we see concentrations of failed waterpoints in Mara and Mwanza regions, which are also located in the northwest next to current project regions. Based on model intepretations, handpump wells installed at communal standpipes with multiple waterpoints have the greatest impact on continued functionality.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%