Predicting the Status of Tanzanian Wells: Technical Report

Project Summary:

The aim of this project was to classify water pumps in Tanzania into the following three categories: “functional”, “non-functional”, and “needs repair”. After cleaning and processing the data, several models were utilized to classify the pumps into the three categories. The highest performance was achieved with a Random Forest Model which predicted the water pump categories with approximately 80% accuracy.

Background:

In many parts of Tanzania, women are responsible for the collection and provision of water for their households. Water sources are often far from their homes, so water collection is labor intensive and time consuming. The responsibility also often falls on girls or younger women who often have to compromise their education to get clean water.

Additionally, the quality of the available drinking water is often of a poor standard and can be unsafe to drink. The quantity of water in some villages is also very seasonally dependent with insufficient availability for much of the year.

The Question: Why model water pumps in Tanzania?

Many non profit and government organizations have sought to address these issues by constructing water pumps around Tanzania, but there have been far fewer efforts to maintain the pumps or see if they are still working. In an effort to use data to inform water policy, The Tanzanian government partnered with DrivenData to create a data science challenge to better identify important predictors of a functional well. The Tanzanian Ministry of Water will use the results from the challenge to classify wells into the three aforementioned categories: "functional", "non-functional", and "needs repair".

The Data:

The data set was provided by DrivenData. It was collected by the Taarifa non-profit and the Tanzanian Ministry of Water. It contains information covering around 30 features and 60,000 wells.

Assumptions about the data include the assumption that the data was collected properly and uniformly in a way that did not introduce bias into the data set. Also there is the assumption that the data is accurate.

Cleaning and Transforming the Data:

There were several features that clearly did not have much value in modeling that were dropped. Some of these features included ones that were nearly all zeroes (which appeared to indicate the true value was not recorded), were largely null, or had no predictive value and were almost entirely comprised of unique values (e.g. “ID”)

The missing “construction year” values were imputed using the mean.

The remaining null values were filled in using fillna.

Clustering Based on Latitude/Longitude:

I used the DBSCAN, density-based clustering algorithm to create groups based on latitudes and longitudes. The DBSCAN resulted in 13 clusters which the rows were grouped into. I then added these clusters as a feature to the training and testing data.

Heat map:

I generated a heat map to look at the correlation between the numerical variables.

Adjusting Categorical Variables:

Twenty-seven of the features were categorical which was problematic for modeling purposes. In order to make these features usable for modeling process they needed to be changed to numerical form. I did this by using get-dummies on the object value feature columns. In order to maintain consistency between the test set and train sets I subtracted the column values of one from another and filled in the remaining column values with zeroes.

The y_train values also were in categorical form, meaning they also needed to be transformed to numerical values. I wrote a function that transformed the values from “functional”, “non functional”, and “functional needs repair” to 3, 2, and 1 respectively.

Calculating Baseline for Models:

By calculating the value_count of each of the three y-values in y_test and dividing it by the length of y_test, I was able to calculate a baseline accuracy score to compare my models to. It was 54%.

Modeling:

I used several different models with and without PCA and with and without ADABoostClassifier including Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier, SVM, K-Nearest Neighbors, and also used a Neural Network. The Random Forest Classifier with ADABoost had the best performance with an accuracy score of 79%.

Discussion / Interpretation of Results:

The results were indicative of an imbalanced multi-class classification. The majority of the water pumps were classified as “functional”. The most predictive features were latitude and longitude suggesting that where the pump is located is important in determining the status of the pump.

Conclusion:

Predicting which wells need to be repaired or are non-functional can be valuable in resource allocation as the government decides how to tackle water security in Tanzania. With the information from the DataDriven Challenge, future efforts to gather information on the status of the wells will also be better informed.