Gabe Walker
April 11, 2019

A Stacked Ensemble Model to Classify Potable Water Sources in Tanzania

INTRODUCTION

The ultimate goal of this algorithm is to predict the functionality of potable water sources across Tanzania. To do this, it draws on more than 50,000 observations of water sources classified as functional, nonfunctional, or in need of repairs. The observations cover twenty-one regions, comprise seven water source types, and include water sources built over more than forty years. Overall there are forty-one variables associated with each source that serve as predictors for the source's functionality. With this data, from Taarifa, and a powerful algorithm, the Tanzanian Ministry of Water can better plan future maintenance and address drinking water shortages around the country.

The method I employ is a "stacked ensemble" machine learning algorithm. It is stacked because it combines two levels of modeling, and an ensemble because it makes use of multiple types of models in concert. This approach has proven effectiveness for a wide variety of applications and is especially known for its accuracy in machine learning competitions. It combines predesigned machine learning packages in a novel way for this particular challenge.

MODELING APPROACH

The model-building begins with data cleaning and processing. First, I drop certain features that provide no predictive information, including site ID (since they all differ across observations) and the "recorded by" field, which is GeoData Consultants, Ltd., for all observations. I also remove the ward name and main geographic marker for the random forest model because they contain too many levels for the model to process. I next impute missing data with a method known as Fully Conditional Specification, an iterative process that uses non-missing values to guess missing ones. In the raw data, the permit and public meeting status include many missing values. I also take the additional step of imputing missing years of construction, many of which (over 20,000) are coded as 0. With this additional step, construction year becomes one of the most important predictors (at least in the random forest model). This imputation of missing values ideally gives the later models greater predictive power.

Next I employ three models for classifying the functionality of potable water sources on a training subset of the known data:

- **Random forest**, a stochastic method that builds many small decision trees and aggregates their results to provide a best guess for each observation.
- ***k*-nearest-neighbor** (KNN) with "leave one out" cross-validation, a nonparametric approach that classifies each observation (one out-of-fold point) based on the most closely associated points for all the other ($n - 1$) observations.
- **Multinomial logistic regression**, a type of regression for classification that can predict more than two classes of outcomes (as regular logistic regression does).

In training the models I found that including all the features provided the highest (best) F1 score, a measure of a model's false positives and false negatives, for the random forest. Thus, I included all the features for the other two models as well. A k value of 5 gave the highest F1 score for the

KNN model. I also used 100 trees in the random forest model since a greater number did not increase overall accuracy.

Finally, I "stacked" the three base models by training another multinomial logistic regression on their predicted values. In simple terms:

$$y_{actual\ outcomes} = x_{model\ 1\ predictions} + x_{model\ 2\ pred.} + x_{model\ 3\ pred.}$$

This new model should essentially learn which of the three lower models provides the best guess of the actual water source status in the training data. Then I use this final model on our base model predictions for the *test* data for which we want to make predictions. An ideal stacking model combines the results of different kinds of base-level learners. Because random forest, KNN, and logistic regression use different mathematical approaches, their combined results should have stronger predictive power overall.

FINAL MODEL

On a training subset of around 40,000 observations, the model achieved an F1 score of around .84 (where 1 is completely accurate predictions). Because our final model learns from the entire dataset of known water sources, we expect the score for out-of-fold predictions to be slightly higher. The model also runs quickly, taking under 10 minutes for missing data imputation, three base models, and classification of 2,000 observation points.

It should be noted that this model is a first attempt. Although its accuracy is satisfactory, future adjustments may boost its predictive power. These include using feature selection (or even multiple feature sets), including more and more types of base models, and including more levels of top-level learners. Additionally, at least for the training case, the F1 score of the stacked model was no different than that of the random forest alone, so it is unclear whether stacking in this context actually adds to predictive power.

BIAS AND TRANSPARENCY

A model can only be as good as its training data. The raw data from Taarifa is skewed in some dimensions, such as overrepresenting certain geographic regions (2,000 observations for Njombe and fewer than 500 for the least-represented 80 regions) and very small populations (median of 25 people around a water source but maximum of 30,000 in Masuguru, Tanga). It is also possible that the surveyors recorded some inaccuracies during data collection that we cannot readily observe from the training data.

An advantage of this model is its accuracy. But because of its complexity, policymakers without an understanding of statistical methods or machine learning may not grasp how it works. The model aims to maximize performance, since correct classifications are what ultimately matter for the Tanzanian Ministry of Water and the citizens it serves. As long as policymakers understand that all models are imperfect representations of reality, and that this one will often—but not always—be accurate, it can serve as a useful guide for increasing potable water access across the country.