Tanzanian Water Well Classification

Predicting the Functionality of Water Pumps in Tanzania

By Andrew Muller

Overview

- Tanzania is struggling to get clean water to most of its population
    - 50% of households have basic water coverage
    - 24% of households have sanitation coverage
    - supply is intermittent in 17 of 20 urban areas

- Data provided by an active competition on drivendata.org
    - full data on 59,400 water pumps provided
    - data without target variable provided for 14,850 pumps for contest submission
    - 39 independent variables
    - ternary (three-class) classification problem
        - functional
        - functional needs repair
        - non functional

Overview (cont)

- Sources:
  - https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/
  - https://en.wikipedia.org/wiki/Water_supply_and_sanitation_in_Tanzania

- Libraries Used:
  - imblearn
  - matplotlib
  - numpy
  - pandas
  - sklearn
  - xgboost

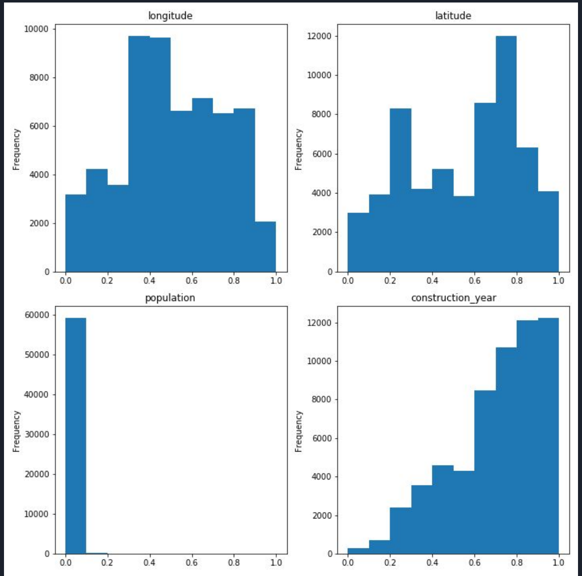
Exploratory Data Analysis

Categorical Variables

- Removed 22 variables that were copies of others or otherwise useless
- Many categorical variables have thousands of categories
- All categories with less than 1% representation go into an "other" category
- Created dummies of categorical variables for a total of 100 features

Continuous Variables

- Scaled all values with MinMaxScaler
- Imputed missing values with KNN imputation



Models

Baseline Models:

- Random Forest: 79.30%
- Bagged Trees: 78.38%
- K Nearest Neighbors: 77.98%
- Support Vector Machines: 77.04%
- Gradient Boost: 74.99%
- Decision Tree: 74.64%
- XGBoost: 74.42%
- Logistic Regression: 73.40%
- Adaboost: 72.74%
- Naive Bayes: 54.23%

After Hyperparameter Tuning:

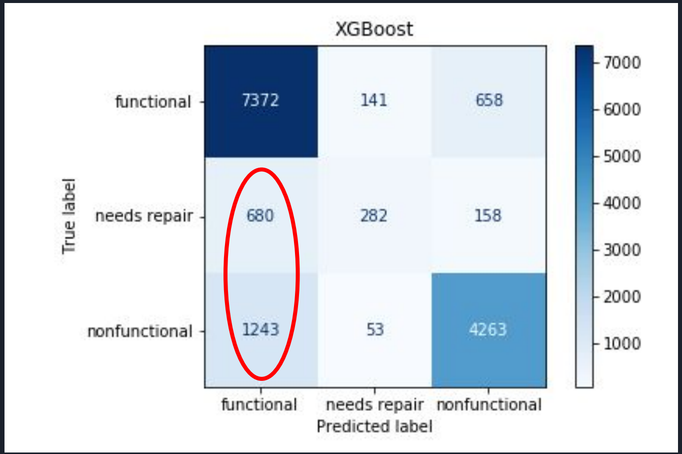
- XGBoost: 80.25%
    - improvement of 5.83
- Random Forest: 80.06%
    - improvement of 0.76
- K Nearest Neighbors: 78.57%
    - improvement of 0.59
- Support Vector Machines: 77.70%
    - improvement of 0.66

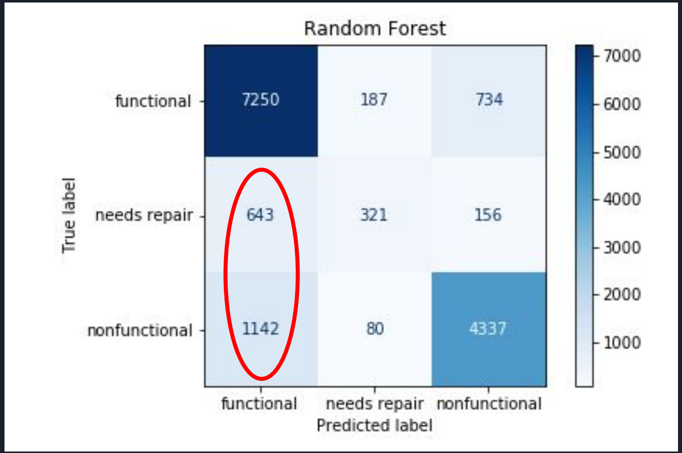
Models (cont)

Models Submitted to Competition:

- XGBoost: 81.50%
- Random Forest: 81.49%
- K Nearest Neighbors: 79.99%
- Support Vector Machines: 78.04%

First place is currently 82.94%





Conclusions

- XGBoost performed the best in the competition
- Random Forest performed nearly as well, with much fewer false positives
- Accuracy may not be the best performance metric for these models
- Use Random Forest model instead of XGBoost

Future Improvements

- Fix accidental data leakage when imputing data before train-test split
- Remove fewer categories at the cost of processing time
- Run larger grid searches on hyperparameters
- Check confusion matrix of submission data if it can be provided

Thank You

https://github.com/MullerAC/tanzanian-water-well-classification

Andrew Muller

Any Questions?