HarvardX: PH125.9x Data Science

Khushnud Sapaev

January 5, 2019

CLASSIFICATION ANALYSIS OF WATER PUMPS FUNCTIONALITY IN TANZANIA

INTRODUCTION

Water is a basic necessity and right for all humans, but not all populations in the world have access to clean and potable water. Quality of water delivered and used in households has an essential influence on hygiene and public health (Howard et al., 2003). Effective and sustainable management of water resources is a fundamental factor for guaranteeing sustainable development (Dungumaro & Madulu, 2003). Water is a public resource fundamental to life and in nurturing the environment and plays a dominant position in the social and economic development of the country (maji.go.tz). Tanzanian government started actively becoming involved in the construction of rural and urban water supplies in the 1950s to grant the nation access to drinkable water. The government's share for water development projects covered 75% of the capital cost, the remaining 25% was contributed by the local authorities (Maganga et al., 2002).

Major freshwater sources in Tanzania include lakes, rivers, streams, dams, and groundwater (maji.go.tz). However, they are not well distributed all over the republic. Several areas are deficient in both surface and groundwater sources. The sustainability of the population and environment in Tanzania depends mainly on proper water resources management. In the 1980s Tanzania adopted a River Basin Management Approach for water resource management by dividing the country into nine basins. In 1991 the first National Water Policy was launched to expand the changes in the water segment (Sokile & Koppen, 2005). In 1995 Tanzania's water resources policies and institutions were reviewed by the Government of Tanzania, World Bank, and DANIDA (Sokile & Koppen, 2005). Tanzania's National Water Policy is separated into the following legislation: the Water Resources Development Act, the Rural Water Supply Act, and the Urban Water Supply Act (Sokile & Koppen, 2005).

OBJECTIVE

The mission of Tanzanian Ministry of Water is to ensure the protection of water sources for Tanzanians and community involvement for sustainable development and management of water resources (maji.go.tz). Availability of adequate water supply of good quality reduces time spent in fetching water, increases health standards, and reduces the prevalence of water-borne diseases such as diarrhea and cholera.

Modeling the functionality of waterpoint pumps based on environmental and geographic locations would be helpful to target the quality of water and the installment of different types of pumps, and to systematically supply the population with drinkable water. Therefore, the purpose of this study is to explore the relationship between the functionality of installed pumps and installation, management, quality, quantity, as well as geographic, seasonal, and other factors in terms of data mining algorithms.

Software R is used to predict the pump operation functionality via classification algorithms. Support Vector Machines (SVMs), K-Nearest Neighbors (kNN), Gradient Boosting Machines, Extreme Gradient Boosting, and Random Forest classification algorithms are chosen to classify the functional status.

DATA DESCRIPTION

The data for the study is provided by the Tanzanian Ministry of Water and available at "DrivenData.org" under the data science competitions section. The datasets include training, the label of the pump functionality status, and test datasets of installed water pumps in Tanzania. Each water pump has a unique identification number and labeled as one of the following functionality positions: “functional”, “functional but needs repair”, and “non-functional”. Training dataset and label dataset have *59,400 observations* of *40 and 2 variables* respectively. Test dataset consists of *14,850 observations* of *40 variables*. Out of the 40 features in the datasets, 31 are categorical variables, 7 are numerical variables, and 2 are date variables. Training dataset and label dataset have the common column: the ID of the pump. Data is collected from geographical locations, waterpoint type, source of water, water quantity, water quality, extraction of water, basin type, construction year, water pump installation, and pump operation details. Both training and test datasets have “zero valued” observations that represent missing values. The purpose of the study is to predict the status of water pumps in the test dataset based on the model used in training data.

DATA PREPARATION

a) Combining training dataset and label dataset: Both datasets have the identical ID attribute, therefore datasets were combined by ID of the water pump.

b) Removing duplicate or unnecessary columns: First, variables with similar content were checked to keep one of the variables from the group. Second, variables were checked for missing data to remove columns with a large number of missing observations. A total of 22 columns were removed from both training and test dataset including ID variable to prevent the over-fitting problem.

c) Recoding variable “installer”: The table of most frequently appearing installers showed that “government” and “central community” appear in different formats, also some installers consist of three letters or more. I decided to lower all letters and trim to three strings; then I kept the 25 most frequent installers and the rest of the installers labeled as others.

d) Replacing zero valued “construction_year” values to median values based on installer: For each installer, the zero-construction year was replaced with the median.

e) Categorizing “lga” column into three categories: “rural,” “urban,” and “other”.

f) Creating the season variable: I extracted months from “date_recorded” variable and grouped months by seasons in Tanzania.

g) Creating the age of the pump subtracting “construction_year” from the maximum “date_recorded” year.

h) Renaming blank factor level of scheme management variable as “None” to match factor levels of both training and test datasets.

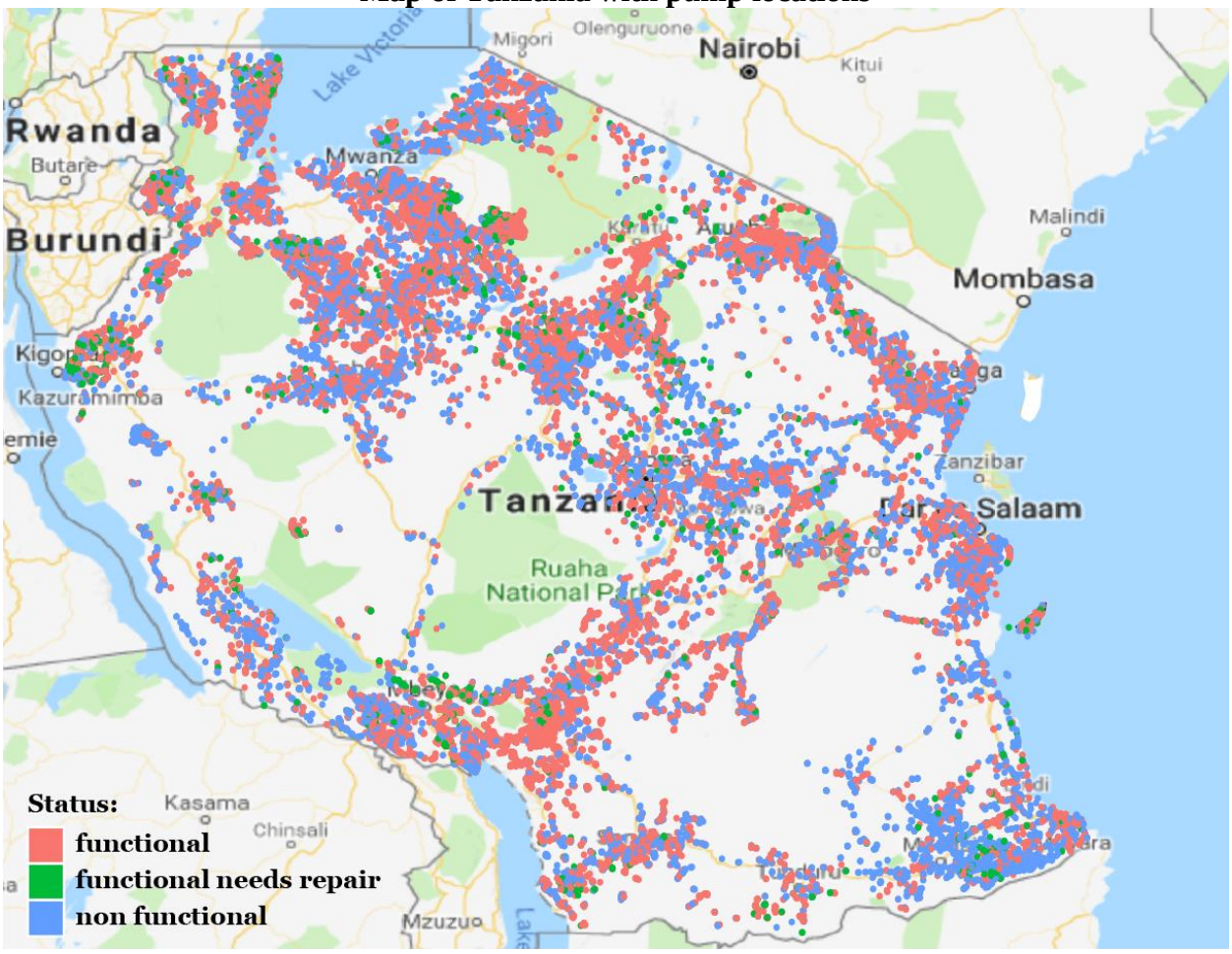
A popular way to procedure data for classifiers is to split the training dataset into two sets, a training set, and a test set. The classification model will be built on the training set and applied to the test set. The test set is unseen by the model, so the resulting performance will be a decent

prediction to what will be seen when the model is applied to unseen data. For this project, the dataset was randomly split at 0.7 and 0.3 ratio: ***training set*** – 41,581 observations of 19 variables; ***test set*** – 17,819 observations of 19 variables.

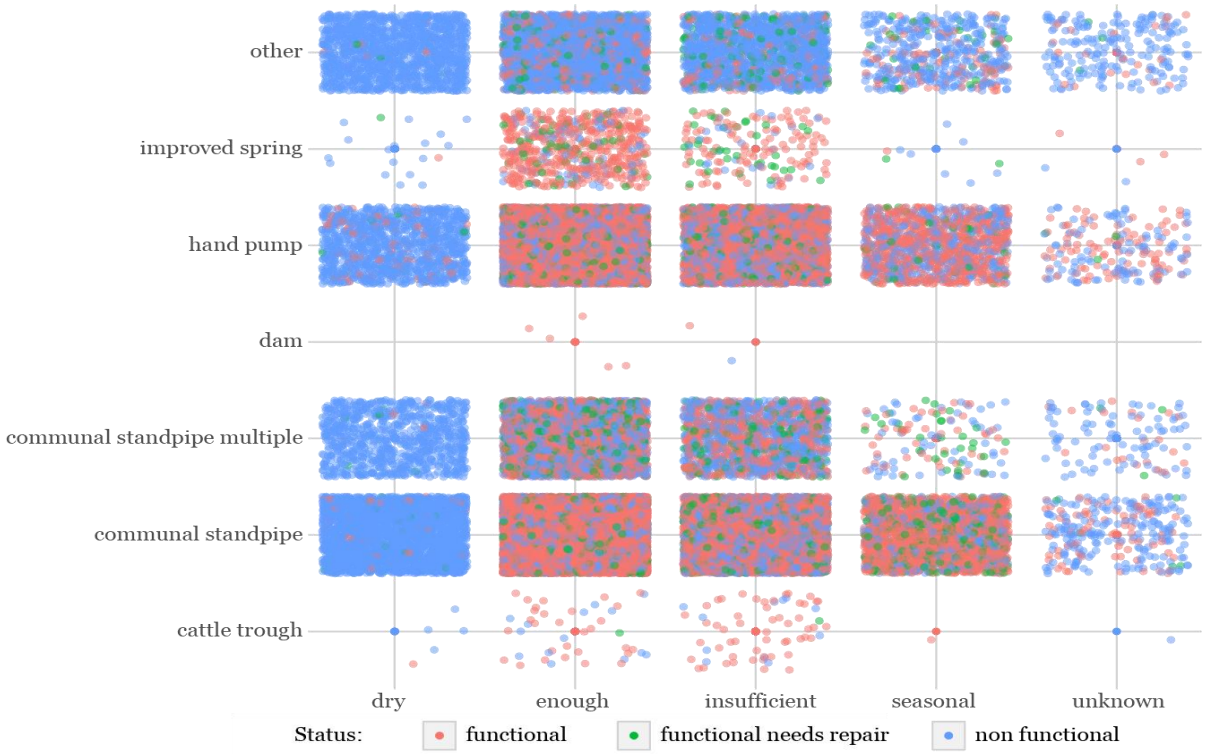
DATA VISUALIZATION

Data visualizations were made to provide insightful details of the observations, including distribution of the pumps across Tanzania.

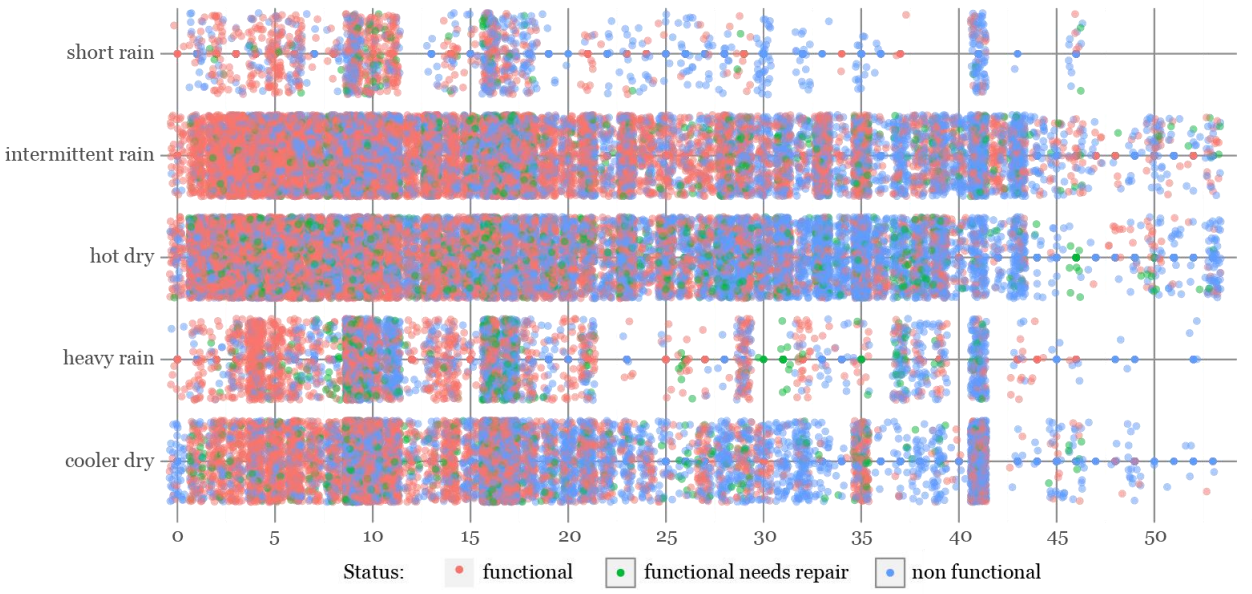
Map of Tanzania with pump locations



Status of pumps in Tanzania by water quantity and waterpoint type



Distribution of pumps in Tanzania by age and season



CLASSIFICATION RESULTS

SVM Linear. Support Vector Machine uses a kernel technique to transform the data and then based on these transformations it finds an optimal boundary between the possible outputs. Three-fold cross-validated SVM algorithm was run on sample sizes: 27,721; 27,721; 27,720. The accuracy of the model on training data was 0.7294437, kappa 0.4587456.

	Precision	**Recall**	**F1**
Class: functional	0.6975437	0.9361372	0.7994176
Class: functional needs repair	NA	0.0000000	NA
Class: non-functional	0.8365066	0.5903315	0.6921825

Table 1. Precision and Recall of three-fold cross-validated SVM linear algorithm

Stochastic GB. Gradient boosting constructs additive regression models by sequentially fitting a simple parameterized function (base learner) to current "pseudo"-residuals by least squares at each iteration. The model predicted the best accuracy with the following parameters: n.trees = 150, shrinkage = 0.1 and n.minobsinnode = 10.

	Precision	**Recall**	**F1**
Class: functional	0.734247	0.9079260	0.8119022
Class: functional needs repair	0.600000	0.1413127	0.2287500
Class: non functional	0.815429	0.6607273	0.7299718

Table 2. Precision and Recall of three-fold cross validated Stochastic GB algorithm

K-NN. K-Nearest Neighbors is robust to noisy training data and is effective in case of large number of training examples. Three-fold cross validated k-Nearest Neighbors were run at sample sizes: 27,720; 27,720; 27,721. The best accuracy was obtained at 5 nearest neighbors.

	Precision	**Recall**	**F1**
Class: functional	0.7798615	0.8379663	0.8078705
Class: functional needs repair	0.4707207	0.3227799	0.3829592
Class: non functional	0.7716210	0.7362348	0.7535127

Table 3. Precision and Recall of three-fold cross validated kNN algorithm

SVM Radial. Support Vector Machine with Radial kernel is used when the problem is not linearly separable and performs better prediction results in highly dimensional space. The model predicted the best accuracy with the following parameters: sigma = 0.005444161 and C = 1. Summary of sample sizes: 27,720; 27,720; 27,722.

	Precision	**Recall**	**F1**
Class: functional	0.7422740	0.9282836	0.8249231
Class: functional needs repair	0.6151203	0.1382239	0.2257251
Class: non functional	0.8512717	0.6746020	0.7527092

Table 4. Precision and Recall of three-fold cross validated SVM radial algorithm

Random Forest. Random Forest algorithm can handle high dimensional spaces as well as large number of training examples. The algorithm was chosen with the minimum model error with the final value used for the mtry = 62.

	Precision	Recall	F1
Class: functional	0.8081011	0.8720678	0.8388668
Class: functional needs repair	0.5012019	0.3220077	0.3921016
Class: non functional	0.8250306	0.7885205	0.8063625

Table 5. Precision and Recall of Random Forest algorithm

XGBoost. Extreme Gradient Boosting is an implementation of gradient boosted decision trees designed for speed and performance. The model was built on the following final parameters: number of rounds=100, max_depth=10, eta=0.1, gamma=1, colsample_bytree=0.7, min_child_weight=2, and subsample=0.75

	Precision	Recall	F1
Class: functional	0.7877704	0.9106128	0.8447491
Class: functional needs repair	0.6252427	0.2486486	0.3558011
Class: non functional	0.8510951	0.7604790	0.8032395

Table 6. Precision and Recall of three-fold cross-validated eXtreme Gradient Boosting algorithm

MODEL EVALUATION

Cross-validation is used to evaluate the model performance in the train set. For each model, the number of cross-validation folders is fixed to three. All models are qualified in the prediction of pump status. Besides the model evaluation and prediction, the confusion matrix is also an important tool to explore deeper information. Results show that linear separation does not detect the pump status "functional but needs repair". Linear models used in the project, such as SVM Linear could not predict the aforementioned pump status.

Geographic locations, pump age, waterpoint type, water quantity, population, payment type, and water extraction type were the main predictors of pump status (Table 7).

	XGBoost	Random Forest
longitude	100.000	42.986
latitude	87.782	42.41
pump_age	53.737	45.219
gps_height	44.445	30.675
waterpoint_type: other	45.455	25.679
quantity: enough	39.080	70.955
quantity: seasonal	34.058	34.173
population	30.044	29.979
quantity: insufficient	29.078	44.647
extraction_type_group: other	19.495	19.964
payment_type: never pay	14.967	29.626
waterpoint_type: communal standpipe multiple	8.602	26.003

Table 7. Variable importance of XGBoost and Random Forest models in Caret package

After considering the Accuracy score, F-1 score, Precision and Recall values for the sampled test set, Random Forest and Extreme Gradient Boosting Classifiers are performing well (Figure 1). These two models had the highest overall accuracy on submitted test data (80 % and 80.48%) among the other models. The overall accuracy of each model on test data is summarized in Table 8.

#	**Model Name**	**Accuracy** (sampled test set)
1.	SVM Linear	73.58%
2.	Stochastic Gradient Boosting	75.72%
3.	k-Nearest Neighbor	76.14%
4.	SVM Radial	77.34%
5.	**Random Forest**	**80%**
6.	**eXtreme Gradient Boosting**	**80.48%**

Table 8. Model prediction performances.

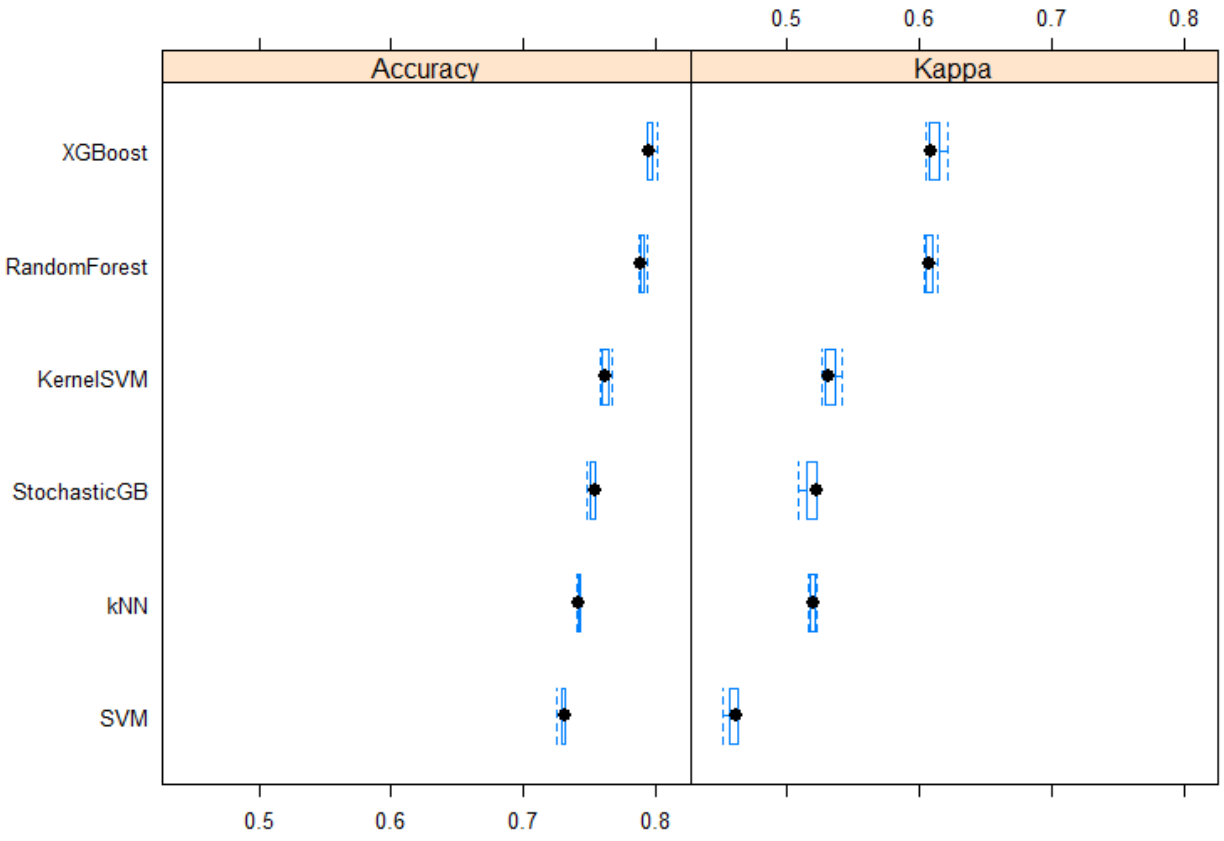


Figure 1. Accuracy distribution of the classification algorithms on training dataset

CONCLUSION

The impact of each variable for pump status varies by classification model. R's Caret package returned the same variable importance for kNN and both SVM Linear and Kernel models. It is important to note that linear models couldn't separate water pumps labeled as "functional but needs repair." Feature importance of the models shows that ***available water quantity, age of the pump, height of the land***, ***payment for water, water extraction type, waterpoint type, and region of the water pump*** are going to influence the functional status of the pumps more than the other predictors.

Predicting the pumps which are "functional but needs repair," decreases the overall cost for the Tanzanian Ministry of Water. Accurate prediction can improve the maintenance processes of the water pumps and verify that fresh, drinkable water is available to communities across Tanzania. Distinguishing pump status, especially between "non-functional" and "functional but needs repair" could help the Tanzanian water industry reduce expenses. Maintaining the pump would cost significantly cheaper than installing a new pump. Moreover, the classification model is useful in the installation of a water pump at a specific location.

REFERENCES

Dungumaro, E. W., & Madulu, N. F. (2003). Public participation in integrated water resources management: the case of Tanzania. *Physics and Chemistry of the Earth, Parts A/B/C*, *28*(20-27), 1009-1014.

Howard, G., Bartram, J., Water, S., & World Health Organization. (2003). Domestic water quantity, service level, and health.

Maganga, F. P., Butterworth, J. A., & Moriarty, P. (2002). Domestic water supply, competition for water resources and IWRM in Tanzania: a review and discussion paper. *Physics and Chemistry of the Earth, Parts A/B/C*, *27*(11-22), 919-926.

Official website of the Ministry of Water of The United Republic of Tanzania (maji.go.tz), retrieved from http://maji.go.tz/pages/mission-statement

Sokile, S & Koppen, Barbara. (2005). Integrated Water Resource Management in Tanzania: Interface between Formal and Informal Institutions.