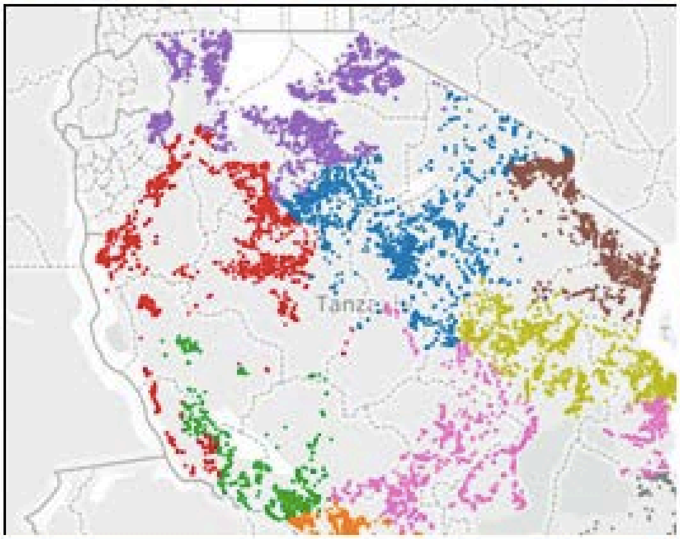
Pump It Up



Predicting the Operating Condition of Tanzanian Wells

Final Executive Summary

Tanzanian Water Ministry

Executive Overview

Throughout the executive summary the group intends to provide very high-level information regarding the completion of the Pump It Up project. In the next few pages you will see information regarding our problem statement and a summary of our approach to exploratory data analysis, data transformation and modeling. The final section of the summary includes our conclusions and recommendations related to the originally agreed upon deliverables and a list of recommended next steps. The deliverables include a final recommended model for predicting the functionality of the wells, a recommendation on which tool should be used to solve future modeling problems and a recommendation on data visualization software and an overview of the data visualizations built to support this report.

Problem Statement

Tanzania's wells are the lynchpin of a healthy and stable society. Without a reliable source of water, local food supplies collapse leading to civil unrest. The Ministry of Water is tasked with ensuring that the 59,400 wells spread across a landmass of 365,756 square miles reliably serve a population of 51.82 million people. With 88% of the water sector being dependent on foreign donors, there is pressure to optimize well maintenance teams and minimize well downtime by deploying teams to proactively maintain wells before they fail.

Approach

Objective

Our project objective is to predict which water pumps in Tanzania are functional, which need some repairs, and which are non-functional. Predictive accuracy will result in two primary benefits:

1) Ensure more reliable access to water in Tanzanian communities
2) Improved efficiency in maintenance operations arising from the ability to proactively identify which water pumps require service prior to failure

Description of Data

The data used for this project is obtained from the *DrivenData* competition website *Pump it Up: Data Mining the Water Table* (https://www.drivendata.org/competitions/7/). Each of the 59,400 records has a unique identification number, 39 attributes to serve as predictor variables, and a response variable.

Data	Description
Data Files	2
Observations	59,400
Fields	32
Categorical Predictors	29
Numeric Predictors	6
Date Value predictor	1
Response Variable Predictor	1 with three levels

Overview of the Data Quality

A review of the data reveals data quality challenges that must be addressed prior to model construction. We observe a material level of missing or erroneous data in multiple predictor variables. We also identify high cardinality (too many levels) in several fields. Finally, missing data is more problematic than originally expected, as a value of zero populates fields for many of the missing values.

Variables with more than 5% of values missing were excluded (with the exception of *Amount of Water* and *Number of Private Pumps*). We excluded additional fields due to redundancy, lack of differentiation between observations, errors and extreme cardinality. Thirty predictors were retained for analysis.

Description of Transformation of Data

Based on our data review, we identified the need to perform three types of transformations; imputations for missing data, transformations to address the issue of high cardinality and data standardization

Data Imputation

We reviewed potential data imputation techniques and identified two viable options – MICE and missForest imputation. To determine which imputation methodology to implement, we performed both the MICE and missForest imputations on the predictor variables, trained a random forest model on each dataset, and evaluated results using a multivariate ROC technique to obtain an AUC (area under the curve) metric. Results were very similar, however, the dataset with missing variables predicted using missForest had slightly more favorable performance. As such, we continued our analysis with the dataset using the missForest imputations.

Technique	R on PC	Microsoft Azure	Accuracy
MICE	Approximately 6 hours	36 seconds	
Decision Trees (Missing Forest)	Approximately 6 hours	N/A	Slightly more favorable than MICE

High Cardinality

We noted in our data overview that a number of potential categorical predictors had a very large number of levels, referred to as high cardinality. To address this challenge we applied a transformation to several of the predictors to combine levels with smaller numbers of observations.

Data Standardization

We also standardized our data, as some of the proposed modeling options are sensitive to scale. For numeric data we centered the values by subtracting the sample mean, and then we scaled the data by dividing by the standard deviation. For categorical predictors we used dummy coding, which creates an indicator for each level of each categorical variable.

Exploratory Data Analysis and Data Visualization

For this report, the exploratory data analysis (EDA) was enhanced through the creation of an EDA dashboard using Tableau business intelligence software. The dashboard compares all predictor variables with the response variable and can be accessed and manipulated by all team members. This allows each team member to perform their own EDA and choose the variables they would like to analyze. The EDA dashboard can be accessed via the link below.

https://public.tableau.com/profile/rob.herold#!/vizhome/PumpItUp-InteractiveEDA/InteractiveEDA

Understanding the geographic location of the wells is very important when it comes to efficiently deploying teams to fix the wells. To further the teams' understanding of the location of the wells as well as investigate the geographic distribution of the wells by *status group,* well maps were created. The dashboard containing the wells maps can be accessed via the link below.

https://public.tableau.com/profile/rob.herold#!/vizhome/PumpItUp-WellsMap/WellMap

Modeling Approach

We constructed multiple models with R, Azure software and ANGOSS software to identify which performed most favorably with regard to accuracy metrics. While ROC area under of curve was evaluated, predictive accuracy of the test data was the main metric used for the comparisons.

R Models

Models constructed and evaluated with R software include Random Forest, Deep Learning Neural Network, Support Vector Machine, Gradient Boosting, Bagging, and Multinomial Regression. The table below displays the results for the most accurate version of each model type. Based on the predictive accuracy for the test data, we identify the Random Forest as the most favorable modeling option.

Model	Accuracy – Test Data
Random Forest	**0.8080**
Gradient Boosting	0.8013
Bagging	0.6964
Deep Learning Neural Network	0.7688
Support Vector Machine	0.7893
Multinomial Regression	0.7477

Azure Models

Four of Azure's built in multi-class models were run using the same training and test data as the other experiments in R. As with most new products, there are some limitations. At the present time, it only has four multi-class classification models. However, Azure does provide machine learning hyper parameter tuning. Azure randomly selects combinations of hyper parameter values, selecting the best outcome. Each model was tested using each and every possible hyper parameter value combination. The table below displays the predictive accuracy for each model. Run times were included as running many permutations can be lengthy. To mirror what would be expected in a cost sensitive production environment, CPU utilization within the cloud was limited to 1 hour.

Model	Run Time (Minutes)	Accuracy – Test Data
Multiclass Decision Forest	**26:56:00**	**0.80156**
Multiclass Decision Jungle	> 60 minutes*	N/A*
Multiclass Neural Network	7:57:00	0.781413
Multiclass Logistic Regression	4:21:00	0.743869
All Experiments Limited to 1 Hour of CPU Time		

ANGOSS Models

Three test models were run in ANGOSS to see how they compare to the overall models that were run in R and Azure. Looking at the table below, it is clear that ANGOSS falls short of both R and Azure when it comes to accuracy. R and Azure afford the freedom to tweak models as seen fit, while the ANGOSS models are not as adaptable.

Model	Accuracy – Test Data
Random Forest	**0.7553**
Bagging	0.7468
Deep Learning Neural Network	0.736

Conclusions and Recommendations

Predictive Model Recommendation: Random Forest Model created using R

For this project a total of 13 different model and modeling tool combinations were evaluated. Six different types of models were constructed using R, four different types of models were constructed using Azure and 3 different types of models were constructed using ANGOSS. All of the models were evaluated based on the predictive accuracy of the test data. Based on this metric, we identify the Random Forest constructed using R as the most favorable modeling option for this data set.

Modeling Tool	Most Accurate Model	Accuracy – Test Data
R	Random Forest	0.808
Azure	Multiclass Decision Forest	0.802
ANGOSS	Random Forest	0.755

Modeling Tool Recommendation: R

Recommending a predictive modeling approach requires that many factors be considered during the planning process. Economics, skill sets, timeliness and accuracy should all be considered, as well as what the data itself naturally lends itself to. Government agencies like the Water Ministry will want a nuanced balance between economics and accuracy. It is largely up to the key stakeholders to determine which of these factors are the most important in creating a predictive model. For example, the team never even entertained the idea of using SAS for analysis due to its costly licensing.

The process that the team ran through shows that R is an economical and powerful tool. This program produced the best results for the modeling process that the team engaged in. There are downfalls however. It does require specialized skill sets if highly accurate results are required, and it also suffers from performance issues when imputing missing data. Imputing missing data can be done through an overnight process, with results being made available when the staff arrives at work in the morning. Being able to wait for an overnight batch process to impute data, can greatly reduce the cost of a solution.

When Azure's built in algorithms are used, accuracy is nearly as high as accuracy in R, dropping less than one percentage point (Best R model = 0.8080 , Best Azure model =0.8016). Azure outperforms R in processing speed, such that imputation of missing data is handled in minutes, contrasted with the overnight processing in R. A drawback to Azure is that it has a limited set of built in machine learning algorithms. This challenge can be mediated, however by embedding R script within Azure to match R's accuracy. The biggest drawback to Azure is its recurring cost model, which runs contrary to Tanzania's Water Ministry's desire to be cost efficient.

ANGOSS requires a license fee and recurring yearly maintenance fees. Like Azure, it has a limited set of machine learning algorithms, requiring the need to include R code to achieve greater accuracy. In our analysis ANGOSS underperformed, especially in comparison to the models created using R. As such, we do not recommend using ANGOSS in the future, as it is costly and does not deliver top accuracy marks.

When considering these approaches, the team currently recommends that the ministry leverage R because of the superior accuracy and absence of recurring costs. Azure should be reconsidered in 18-24 months as it begins to support more optimized built-in machine learning algorithms. As machine learning algorithms become more commoditized in Azure, the ministry should be able to use Azure to solve many business problems in a timely and accurate manner, without relying on highly specialized R skills sets.

Data Visualization Recommendation: Tableau

The team also evaluated and recommends the Tableau visualization tool. With over 59,000 wells, it is very important to be able to visualize well status based on geography or other factors, such as elevation, water table, age of well, etc. Through the use of Tableau the team was able to create several working dashboards that allow the government of Tanzania to accurately see where working wells are, as well as those that are non-functional or in need of repair.

Dashboard Overview

- **Exploratory Well Map** - Used to understand the geographic location of the wells as it relates to the functionality. Used to investigate the geographic distribution of the wells by status group.
- **Interactive EDA** - Enhances EDA by allowing each team member to perform their own EDA and choose the variables they would like to analyze. The dashboard compares all predictor variables with the response variable and can be accessed and manipulated by all team members.
- **Random Forest Model Evaluation Map** - Used to evaluate the recommended predictive model based on geography in order to identify areas that had higher or lower than average prediction success.
- **Well Maintenance Team Deployment Map** - This map is part of the final project deliverable. It can be used by the Tanzanian government to identify the geographic location of wells that are functional but are in need of repair and the non functional wells in order to coordinate well maintenance teams.

Recommendations for Next Steps

Deployment Strategy

While the final predictive model achieves nearly 81% accuracy in the classification of well status, there is still uncertainty in determining where to deploy maintenance teams. With the current model results we group the “non-functional” and “functional needs repair” wells and determine that we achieve 92% accuracy when the decision to send a crew to a well is binary (i.e. the well is functioning versus needing some form of maintenance). As such, we recommend deploying crews to wells that are either classified as “non-functional” or “functional needs repair” in order to proactively restore or maintain water supplies.

Database Improvement

During the course of our analysis we encountered a material proportion of missing data. We recommend that efforts are made to improve data quality in order to more effectively leverage information. First, we recommend that missing values be determined and populated when feasible. When it is not possible to obtain correct values, it would be advantageous to clearly denote which entries are unavailable rather than populating with zeros.

Predictive Model Refresh

As the recommended predictive model is derived from current data sources, we recommend a periodic review of the model to ensure that the algorithm used to predict well status is modified when appropriate. Furthermore, we recommend a review of the Azure software option in 18 – 24 months to determine whether it would be a more suitable option at that time.