Data Mining the Water Pumps: Determining the functionality of Water Pumps in Tanzania using SAS® Enterprise Miner

Indra Kiran Chowdavarapu, Oklahoma State University, StillWater, Ok
Vivek Damodaran Manikandan, Oklahoma State University, StillWater, Ok

ABSTRACT

Accessibility to clean and hygienic drinking Water is a basic luxury every human being deserves. In Tanzania, there are 23 million people who do not have access to safe Water and are forced to walk miles in order to fetch Water for daily needs. The prevailing problem is more of a result of poor maintenance and inefficient functioning of existing infrastructure such as Hand Pumps. To solve the current Water crisis and ensure accessibility to safe Water, there is a need to locate non-functional and functional Pumps that need repair so that they can be repaired or replaced. However, it is highly cost ineffective and impractical to manually inspect the functionality of each Water Pump. The objective of this study is to build a model to predict which Pumps are functional, which needs some repair and which don't work at all. SAS® Bridge for Open Street Map and SAS® VA has been used to illustrate spatial variation of functional Water points at regional level of Tanzania along with other socio economic variables. With the help of data mining methodologies like HP Random Forest, Decision trees, the important factors that contribute to the functioning of Water Pumps are identified in this paper.

INTRODUCTION

The purpose of this project is to build a model to predict the functionality of Water Pumps in Tanzania and illustrate the locations of the non-functional Water Pumps using SAS® Visual Analytics. The classification of Water Pumps using the champion model will help in predicting the functionality of Water Pumps and expedite the maintenance operations that will ensure clean and accessible Water across Tanzania in low cost and in a short period of time.

DATA COLLECTION AND PREPARATION

The datasets used in this paper is obtained from Taarifa Water points dashboard, which aggregates data from the Tanzania Ministry of Water. Two datasets are used in the paper. one data set contains the geographical information of the Water points and the other contains the current status and information of the Water points. The final dataset obtained after merging the above datasets contained the following variables:

Role	Level	Count
ID	Nominal	1
Input	Interval	6
Input	Nominal	31
Target	Nominal	1

Table 1. Variable Summary

The dataset has one variable with the role ID. The ID variable ‘Source ID’ is a unique identifier used for identifying the Water points and it is removed for the analysis. The variables like ‘region_code’, ‘Scheme name’, ‘Scheme management’ are some of the nominal variables with too many levels. There are a total of 7 nominal variables that are of no interest to the analysis. These variables are not considered in building the model.

The final data set consisted of 59,401 observations and 24 variables. The below table enumerates some of the variables:

Variable	Level	Description
amount_tsh	Interval	Total amount of Water available in Water point
Construction year	Nominal	The date the Water Pump was installed
gps_height	Interval	Altitude of the well
longitude	Nominal	GPS coordinate
latitude	Nominal	GPS coordinate
basin	Nominal	Geographic Water basin
subvillage	Nominal	Geographic location
region	Nominal	Geographic location
population	Interval	Population around the well
quantity	Nominal	The quantity of Water
source	Nominal	The source of the Water
Waterpoint_type	Nominal	The kind of Waterpoint
status_group	Nominal	Current Status of Water point

Table 2. Data Dictionary for the dataset

DATA EXPLORATION

Exploratory analysis indicated that most of the records have Water points with Status_group as ‘functional’. The distribution of the functionality of the Water points can be seen below:

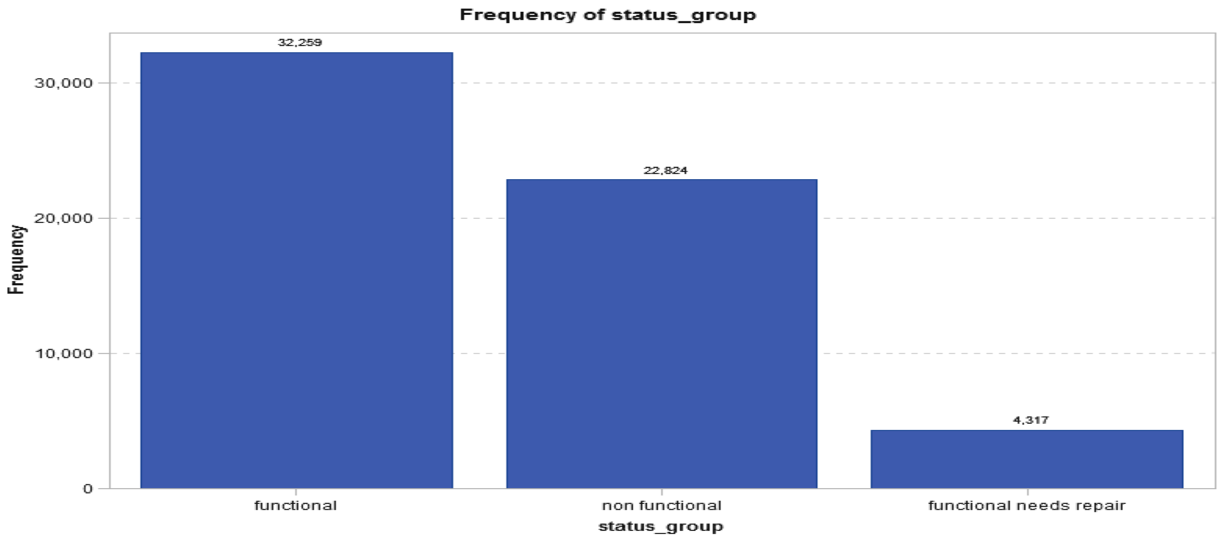


Figure 1. Distribution of Target Variable: Status_group

A total of 32,289 Water Pumps are Functional, 22,824 are non-functional and 4,317 of the Water Pumps needs repair. The regions like kilimanjaro, Shinyanga, Mwanza have the highest percentage of non-functional Water points. The average altitude of these regions are higher than that of other regions.

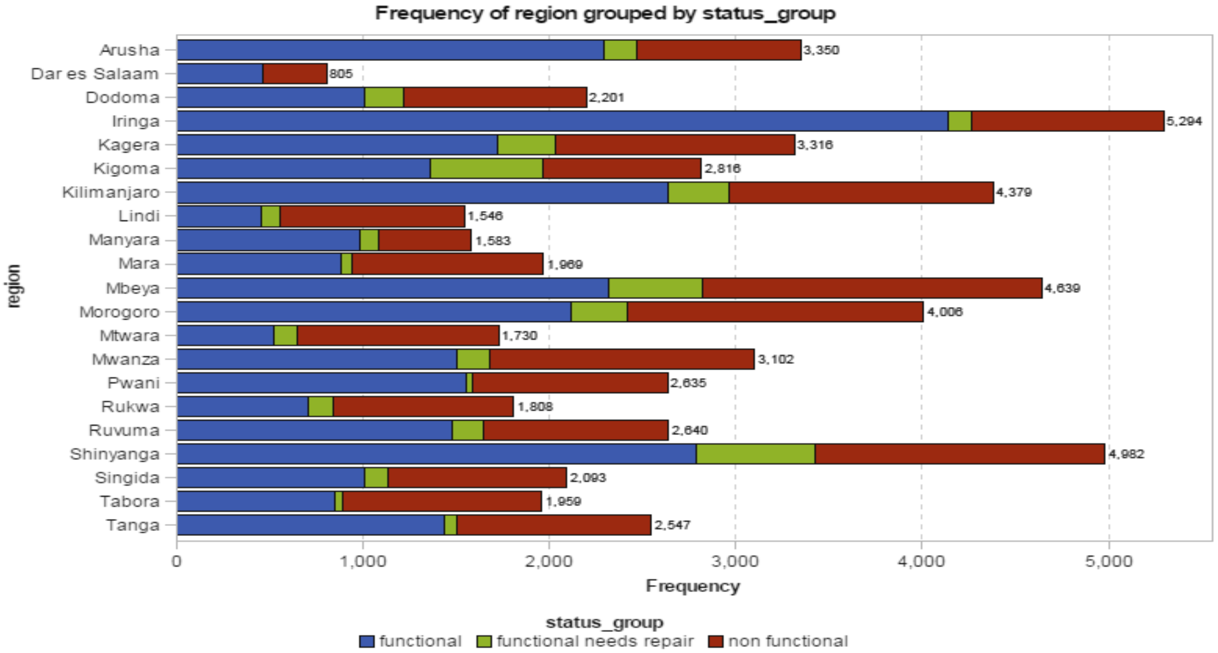


Figure 2. Distribution of Water points by region

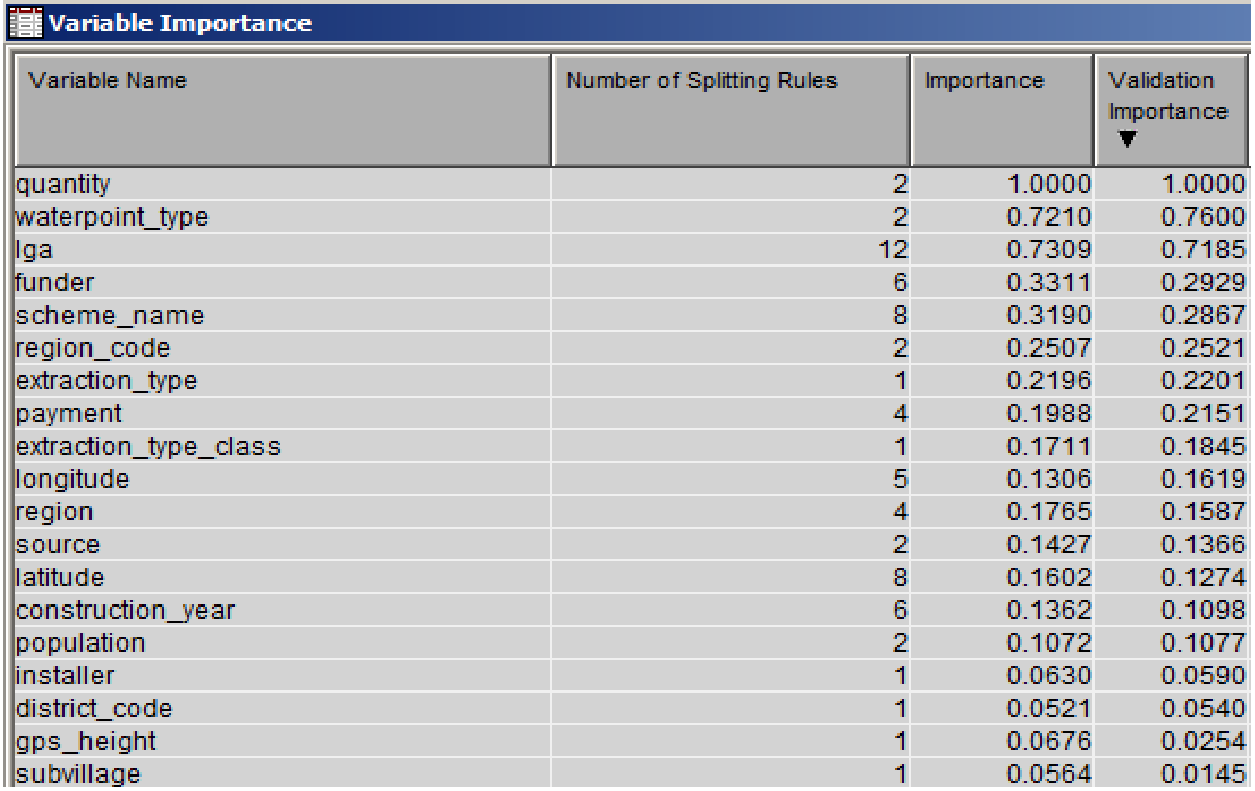
From the data, there are several variables with missing values and incomplete information. The variable amount_tsh indicates the amount of Water in the Water point. There are some Water points with Zero amount of Water and still the functionality of the Water point is indicated as 'functional'. The observations with Zero amount are imputed. All the class variables with missing values are imputed using Tree method and median is used to impute Interval variables.

DATA PARTITION

Data was partitioned into Training data (70%) and Validation data (30%) based on the optimal method of partition ratio, which was required for modeling.

VARIABLE SELECTION

Among all 24 variables, 6 are interval, 5 are binary and 11 are nominal variables. Decision tree is used to find the most important variables that can be used as predictors. Using decision tree, 24 variables were reduced to 19 variables. The important variables that are chosen from decision tree is given below in the Figure 3.



Variable Importance

Variable Name	Number of Splitting Rules	Importance	Validation Importance ▼
quantity	2	1.0000	1.0000
waterpoint_type	2	0.7210	0.7600
lga	12	0.7309	0.7185
funder	6	0.3311	0.2929
scheme_name	8	0.3190	0.2867
region_code	2	0.2507	0.2521
extraction_type	1	0.2196	0.2201
payment	4	0.1988	0.2151
extraction_type_class	1	0.1711	0.1845
longitude	5	0.1306	0.1619
region	4	0.1765	0.1587
source	2	0.1427	0.1366
latitude	8	0.1602	0.1274
construction_year	6	0.1362	0.1098
population	2	0.1072	0.1077
installer	1	0.0630	0.0590
district_code	1	0.0521	0.0540
gps_height	1	0.0676	0.0254
subvillage	1	0.0564	0.0145

Figure 3. Decision Tree for Variable Selection

MODELING

Data mining models such as Decision tree, Neural Network, Multinomial Logistic regression and Random Forest are built using SAS®® Enterprise Miner 14.1. These models were later compared using Model comparison node in order to evaluate the best model using validation misclassification rate as the selection criteria. The model diagram is shown in Figure 4.

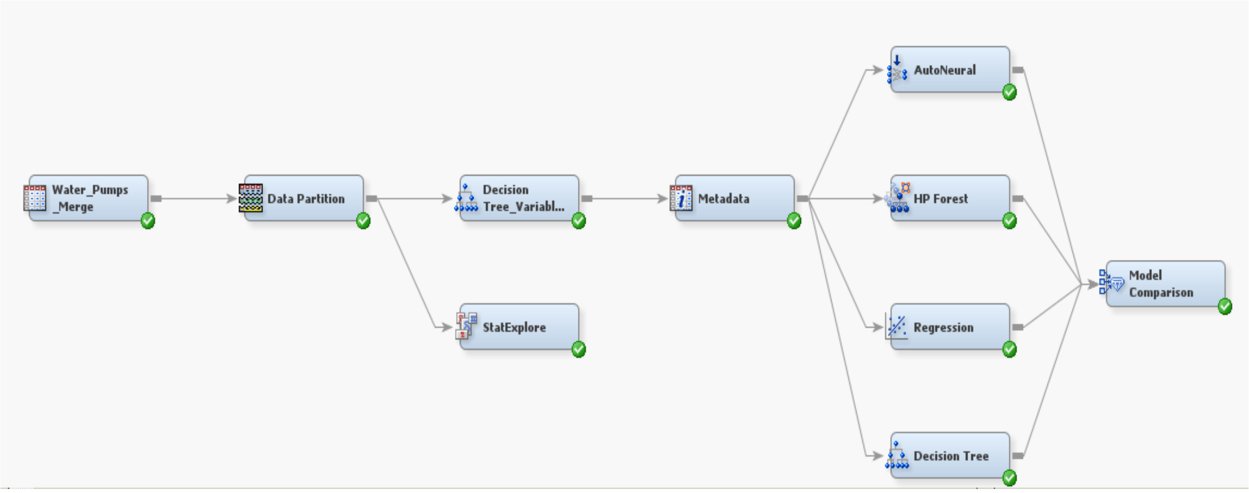
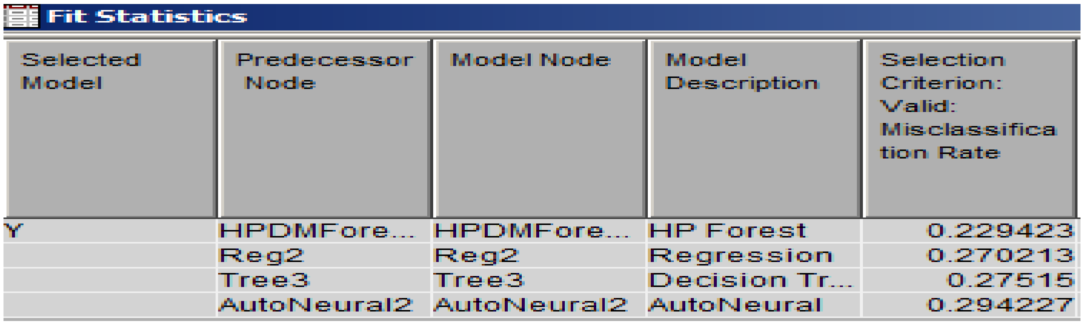


Figure 4.SAS® EM diagram representing the models developed for the study

Figure 5 shows the Fit statistics of Model comparison node with Random Forest model outperforming all the other models with validation misclassification rate of 22.94%.



Selected Model	Predecessor Node	Model Node	Model Description	Selection Criterion: Valid: Misclassification Rate
Y	HPDMFore...	HPDMFore...	HP Forest	0.229423
	Reg2	Reg2	Regression	0.270213
	Tree3	Tree3	Decision Tr...	0.27515
	AutoNeural2	AutoNeural2	AutoNeural	0.294227

Figure 5. Model Selection Fit Statistics

Random Forest Model:

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and can say that the tree "votes" for that class. The forest chooses the classification having the most votes.

Each tree is grown as follows:

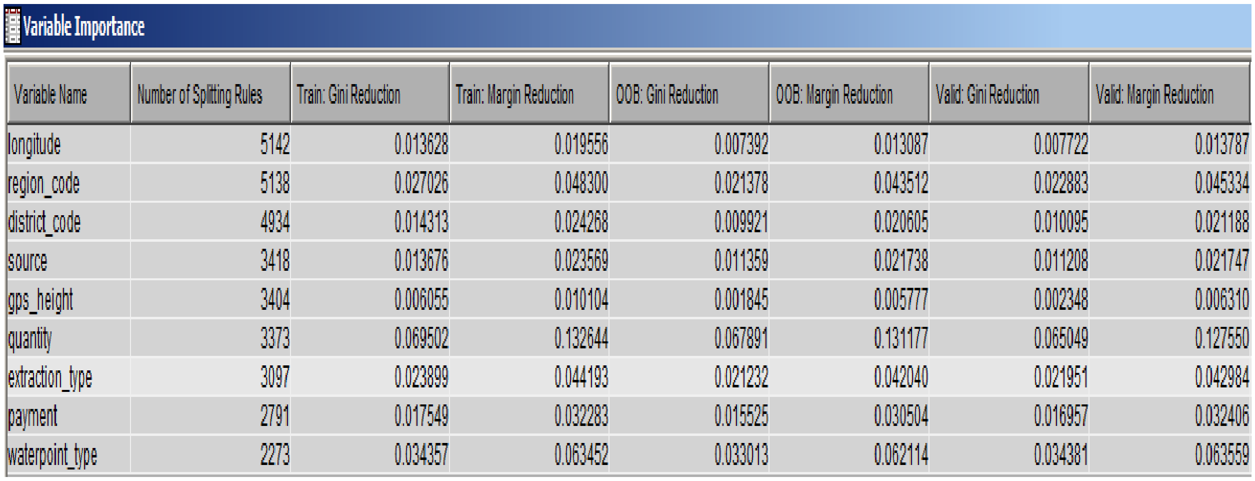
- If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
- If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
- Each tree is grown to the largest extent possible. There is no pruning.

Features of Random Forests

- It is unexcelled in accuracy among current algorithms.
- It runs efficiently on large data bases.
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- It has methods for balancing error in class population unbalanced data sets.
- Generated forests can be saved for future use on other data.
- Prototypes are computed that give information about the relation between the variables and the classification.

- It computes proximities between pairs of cases that can be used in clustering, locating outliers, or (by scaling) give interesting views of the data.
- It offers an experimental method for detecting variable interactions

The significant predictors in the model are longitude, region code, district code, source, gps_height, quantity, extraction type, payment and Water-point type. The variable importance in shown in Figure 6.

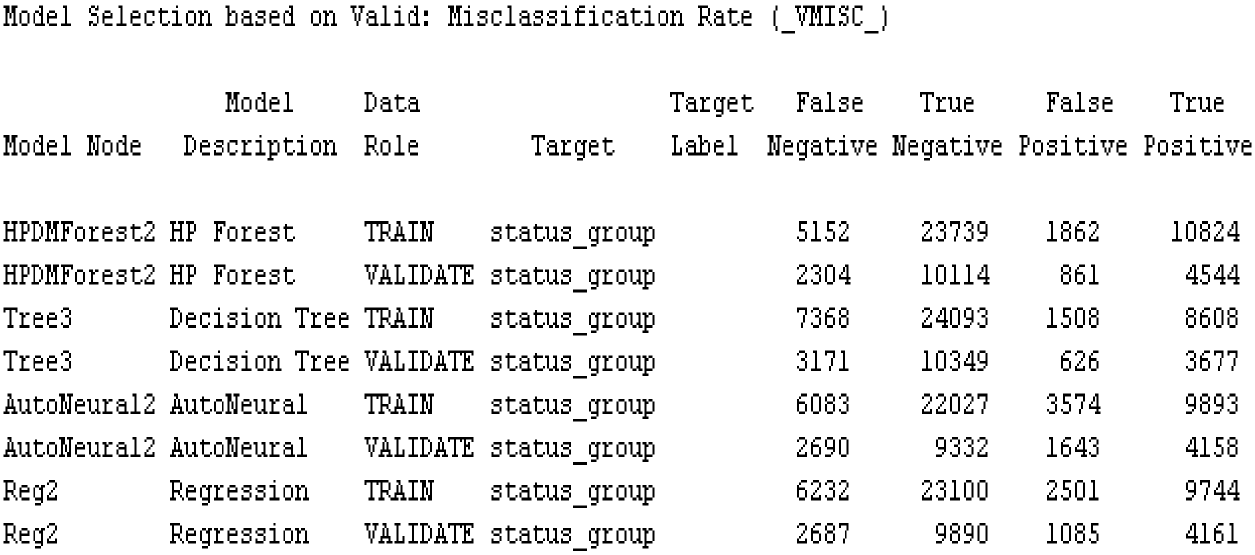


Variable Importance

Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Valid: Gini Reduction	Valid: Margin Reduction
longitude	5142	0.013628	0.019556	0.007392	0.013087	0.007722	0.013787
region_code	5138	0.027026	0.048300	0.021378	0.043512	0.022883	0.045334
district_code	4934	0.014313	0.024268	0.009921	0.020605	0.010095	0.021188
source	3418	0.013676	0.023569	0.011359	0.021738	0.011208	0.021747
gps_height	3404	0.006055	0.010104	0.001845	0.005777	0.002348	0.006310
quantity	3373	0.069502	0.132644	0.067891	0.131177	0.065049	0.127550
extraction_type	3097	0.023899	0.044193	0.021232	0.042040	0.021951	0.042984
payment	2791	0.017549	0.032283	0.015525	0.030504	0.016957	0.032406
waterpoint_type	2273	0.034357	0.063452	0.033013	0.062114	0.034381	0.063559

Figure 6. Variable Importance: Predictors

Confusion matrix given in Figure 7 shows the actual by predicted values for the ternary target variable. **Sensitivity** (or true positive rate is proportion of positives that are correctly identified) from the confusion matrix is 67.75% and **specificity** (or true negative rate is proportion of negatives that are correctly identified) of the model is 92.72%.



```
Model Selection based on Valid: Misclassification Rate (_VMISC_)

                Model      Data                  Target   False    True     False    True
Model Node   Description   Role        Target    Label   Negative Negative Positive Positive

HPDMForest2 HP Forest     TRAIN    status_group            5152     23739    1862    10824
HPDMForest2 HP Forest     VALIDATE status_group            2304     10114     861     4544
Tree3       Decision Tree TRAIN    status_group            7368     24093    1508     8608
Tree3       Decision Tree VALIDATE status_group            3171     10349     626     3677
AutoNeural2 AutoNeural    TRAIN    status_group            6083     22027    3574     9893
AutoNeural2 AutoNeural    VALIDATE status_group            2690      9332    1643     4158
Reg2        Regression    TRAIN    status_group            6232     23100    2501     9744
Reg2        Regression    VALIDATE status_group            2687      9890    1085     4161
```

Figure 7. Confusion Matrix in Validation data

GEO MAP OF WATER POINTS

Geo map option in SAS® Visual Analytics is used to depict the Water points. The latitude and longitude data of Water points are used to locate them visually in a map. The size of the Water point is represented by the size of the population surrounding the Water point area. A snapshot of the Water points can be seen in the Figure 8.

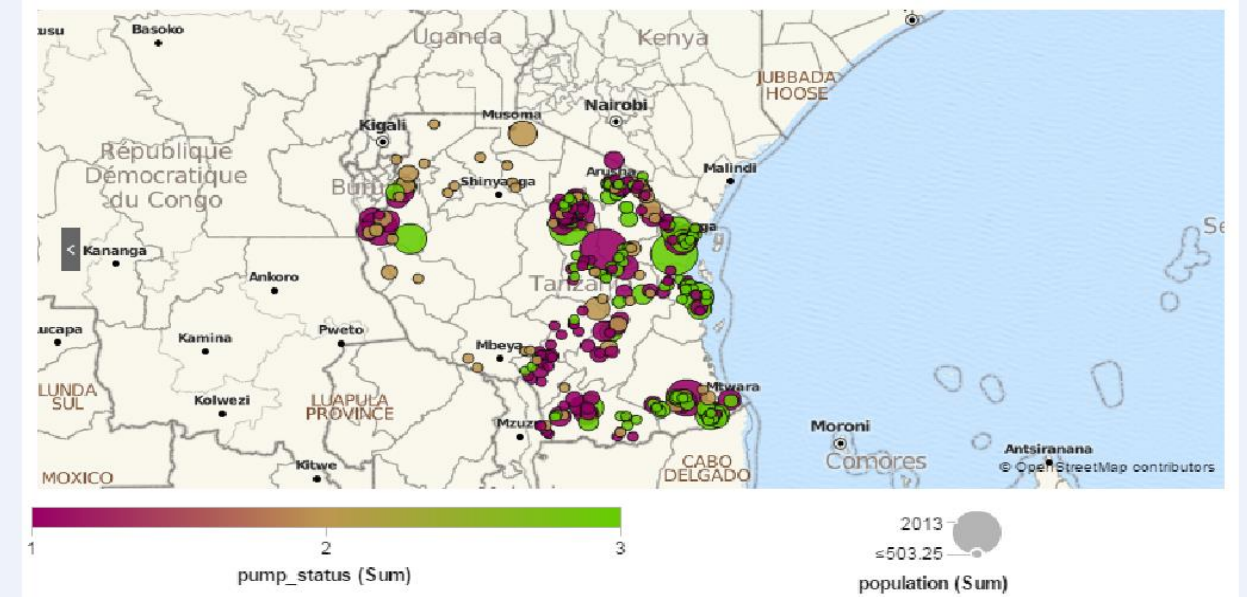


Figure 8. Spatial representation of Water Pumps

In the map, the pink represents the Functional Water points, Green represents the non-functional Water points and Brown represents Water points that needs repair. The representation of Water point locations helps the Tanzanian administration to identify the non-functional Water points and allocate resources to repair the Water points that need immediate attention.

CONCLUSION

Data mining models such as Decision tree, Neural Network, Multinomial Logistic regression and Random Forest were used to predict if a Water pump is functional, non-functional or needs repair. Based on the validation misclassification rate, Random Forest was the champion model in classifying the Water Pumps.

From the model, the major factors that contribute in determining the functionality of the Water pump are identified as are longitude, region code, district code, source, gps height, quantity, extraction type, payment and Water-point type. Also, the sensitivity and specificity of the model were found to be 67.75% and 92.72% respectively.

The average life time of a Water pump is 17.6 years. The life time is found to be impacted by the population within the location of Water pump. The failure rate of Pumps in densely populated areas is 1.6 times that of sparsely populated areas. The Water Pumps towards the eastern region of Tanzania have incurred more maintenance issues than that of Water Pumps in the other regions of the country.

FUTURE WORK

In future, the scope can be extended by overlaying network diagram on the geographical map of non- functional Water Pumps to represent shortest distance between all the Water Pumps. This can be used by Tanzanian government to repair the non- functional units within short time and minimal cost

REFERENCES

- Course notes about Random Forest by Dr. Leo Breiman and Dr. Adele Cutler
  https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

- Anand Chitale and Falho Schulz (2014), "More Than a Map: Location Intelligence with SAS Visual Analytics". Proceedings of the SAS Global Forum 2014.
  More Than a Map: Location Intelligence with SAS® Visual Analytics by Falko Schulz and Anand

- Handbook of Statistical Analysis and Data Mining Applications by Robert Nisbet, John Elder IV and Gary D. Miner
  Statistical Analysis and Data Mining Applications by Robert Nisbet, Gary Miner, Elder John

ACKNOWLEDGMENTS

We thank SCSUG for giving us an opportunity to present our work. We also thank Dr. Goutam Chakraborty and Dr. Miriam McGaugh of Oklahoma State University for their guidance and support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Indra Kiran Chowdaravarpu

Enterprise: Oklahoma State University

Address: Stillwater, OK, 74075

Work Phone: 330-461-7511

E-mail: indra.chowdavarapu@okstate.edu

Name: Vivek Manikandan Damodaran

Enterprise: Oklahoma State University

Address: Stillwater, OK, 74075

Work Phone: 405-762-1880

E-mail: vivekmd@ostatemail.okstate.edu