**DATA MINING AND ANALYSIS**

COURSE PROJECT REPORT

ON

**PUMP IT UP : DATA MINING THE WATER TABLE**

*Under the guidance of*

**Dr. Shankar G**

**Dr. P.G. Sunitha Hiremath**

**TEAM ID**: D08

**TEAM MEMBERS**: Vinayak Dharmatti 268 01FE17BCS242

Veeresh Pattar 286 01FE18BCS431

Abhishek Hiremath 272 01FE16BCS005

Anusha H 277 01FE16BCS036

**Table of Contents**

1. Introduction…………………………………………………………… 3
2. Problem Statement……………………………………………………. 3
3. Related work…………………………………………………………. 3
4. Description of data…………………………………………………… 4
5. Objectives……………………………………………………………. 6
6. Methodology…………………………………………………………. 6
7. Results and discussions………………………………………………. 10
8. Conclusion………………………………………………………….... 11
9. References……………………………………………………………. 11
10. **INTRODUCTION**

Water is essential for the survival of human beings. Many countries like Tanzania has access to a lot of portable water, the countries still face problems where many areas have no reliable access to water. In a household where money is scarce, families have to often spend several hours each day walking to get water from water pumps.

Here we are trying to address the water problem of Tanzania. Water pumps installed in those areas are the main source of water for the population around that area. Even though there were many water pumps constructed with these connections, these water pumps are not well maintained. Many of the water pumps that were built with the donations are now in danger of failing across communities.

We want to help the Tanzanian Ministry of Water in identifying which water pumps are functional, functional needs repairs, and non-functional so that the Ministry can improve the maintenance operations of the water pumps and make sure that clean, potable water is available to communities across Tanzania. By fixing these water pumps early, the people of Tanzania could have improved and continuous access to running water.

In our approach to this problem, we first analyze the data, pre-process it, followed by mining the important data and finally build a model in order to predict the probability and post-process the models to evaluate the best one.

1. **PROBLEM STATEMENT**

In our project we would be trying to address the water problem of Tanzania. So, by predicting the functionality of pumps we could get to the conclusion as to which pumps are working fine, or need repair so that people over there would accordingly manage and make use of the pumps to get proper drinking water.

1. **RELATED WORKS**

One of the solutions to the pump it up challenge was provided by Vaibhav Shukla. He identified the problem with features like gps\_height, population, latitude and longitude which had many missing data points, so he filled those missing data points with the mean and median (as required) of the respective feature in that particular district in which it was lying. He constructed new feature Operational Year from construction year and dropped some irrelevant attributes. He was left with 22 attributes out of 41 after he was done with pre-processing. He also went with Random Forest Classifier with n-estimators =1000 and got public score of 0.8162 and a ranking of 486 out of 5300 contestants.

1. **DESCRIPTION of DATA**

There are three files which will serve as input data namely

“Test set values.csv”(consist 14451 rows and 40 columns)

“Training set labels.csv”(consist 59401 rows and 2 columns)

“Training set values.csv”(consist 59401 rows and 40 columns)

Table 1 : Description of training and testing dataset

|  |  |
| --- | --- |
| **Field** | **Description** |
| amount\_tsh | Total static head (amount water available to waterpoint) |
| date\_recorded | The date the row was entered |
| funder | Who funded the well |
| gps\_height | Altitude of the well |
| installer | Organization that installed the well |
| longitude | GPS coordinate |
| latitude | GPS coordinate |
| wpt\_name | Name of the waterpoint if there is one |
| num\_private |  |
| basin | Geographic water basin |
| subvillage | Geographic location |
| region | Geographic location |
| region\_code | Geographic location (coded) |
| district\_code | Geographic location (coded) |
| lga | Geographic location |
| ward | Geographic location |
| population | Population around the well |
| public\_meeting | True/False |
| recorded\_by | Group entering this row of data |
| scheme\_management | Who operates the water point |
| scheme\_name | Who operates the water point |
| permit | If the water point is permitted |
| construction\_year | Year the water point was constructed |
| extraction\_type | The kind of extraction the water point uses |
| extraction\_type\_group | The kind of extraction the water point uses |
| extraction\_type\_class | The kind of extraction the water point uses |
| management | How the water point is managed |
| management\_group | How the water point is managed |
| payment | What the water costs |
| payment\_type | What the water costs |
| water\_quality | The quality of the water |
| quality\_group | The quality of the water |
| quantity | The quantity of water |
| quantity\_group | The quantity of water |
| source | The source of the water |
| source\_type | The source of the water |
| source\_class | The source of the water |
| waterpoint\_type | The kind of waterpoint |
| waterpoint\_type\_group | The kind of waterpoint |

Table 2 : Training set label description

|  |  |
| --- | --- |
| **Field** | **Description** |
| Id | Unique number given to pump |
| Status\_group | Functionality of the pump |

1. **OBJECTIVES**

* To analyse the given data.
* To pre-process the analysed data.
* To select best suited model for the dataset
* To build model and evaluate the best one

1. **METHODOLOGY**

**Pre-processing:**

Our first approach was to find those features which would be important in predicting the result

That is: Location

Population

Waterpoint

Some attributes have missing values, so they are filled with Nan values.

**Retrieving relevant attributes:**

data.isnull().sum

data.population.min()

data[‘gps height’].replace(0.0,np.nan,inplace=True)

data[‘population’].replace(0.0, np.nan,inplace=True)

data[‘amount\_tsh’].replace(0.0,np.nan,inplace=True)

data. isnull().sum()

**Data visualization**:

* **Integration of our Datasets:**

data\_labels=pd.read\_csv(“rC:\User\Anusha\OneDrive\Desktop\dma csv\training set values.csv”)

data\_values=pd.read\_csv(r”C:\Users\Anusha\OneDrive\Desktop\dma csv\training set labels.csv”)

Here we integrating both the dataset based on ‘ID’

data=data\_values.merge(data\_labels,on=’id’)

* Data Cleaning:

Features like public meeting , permit, scheme name, payment, quality, source type, water point type, ward, installer are dropped as they didn’t seem to have impact on preceding of status.

In the end of data-processing we are left with 22 features out of 39.

* Data Reduction:

Only top 10 classes of attributes funder and installer has been retained and rest of the classes of those attributes are labelled as ‘other’.

* Data Transformation:

New attributes such as operational\_time, recorded\_month, recorded\_day and recorded\_year were constructed from existing attributes.

All the categorical attributes were label encoded in order to convert them into numerical data.

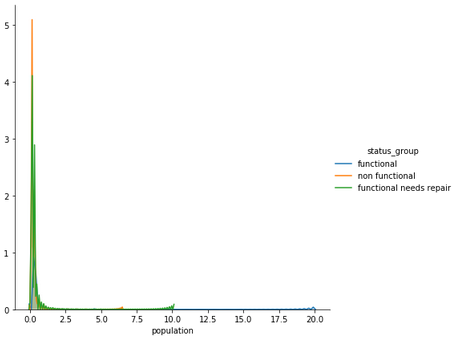


Figure 1 status\_group vs population

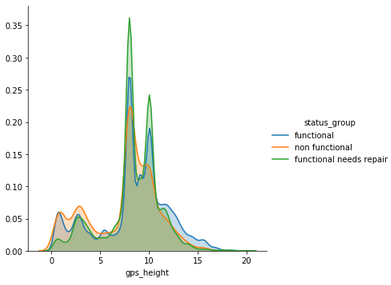


Figure 2 gps\_hesight vs status\_group

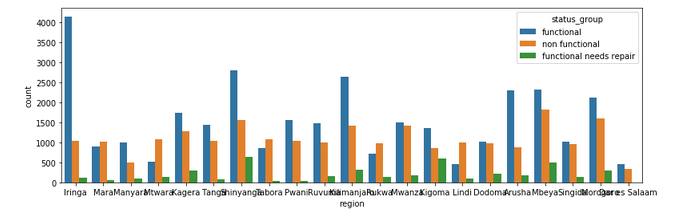


Figure 3 region vs status\_group

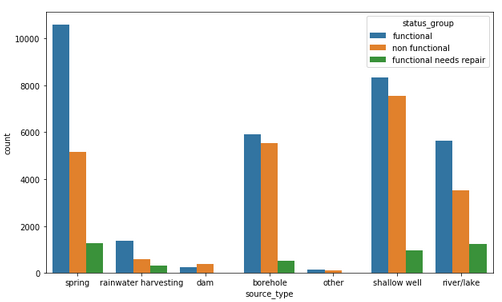


Figure 4 source\_type vs status\_group

* 1. **Building Learning Model:**

After pre-processing, now we are in a stage to apply appropriate algorithm which will learn from training dataset and classifies new samples as functional, non-functional or functional needs repair.

.

Classification models like Linear SVC, XGBOOST, Random forest classifier are used to predict the functionality of the water point.

Table 3: Accuracy of different classifiers

|  |  |
| --- | --- |
| Classifier | Accuracy |
| Linear SVC | 68.34% |
| XGBOOST | 80.1% |
| Random Forest classifier | 81.48% |

1. **RESULTS and DISCUSSIONS**

The Random forest classifier gave us the highest accuracy of 80.1% among all the above-mentioned classifiers.

Accuracy and classification rate were reduced as we tried removing few attributes which we felt weren’t contributing in prediction of functionality.

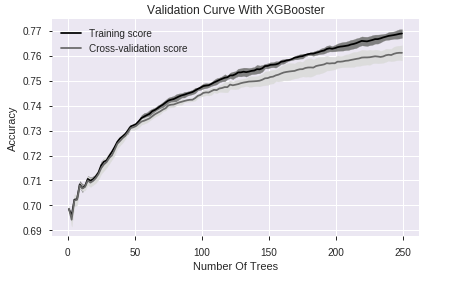


Figure 5 Validation curve with XGBOOST

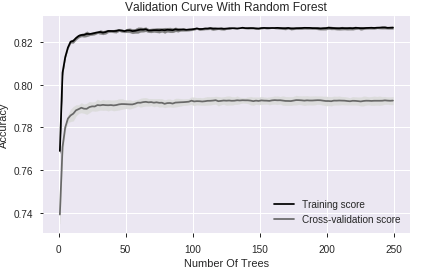


Figure 6 Validation curve with Random forest

1. **CONCLUSION**

Hence, after trying various approaches to solve this problem we could get an insight on how predictive models behave for prediction problems.

After predicting the functionality, it was found that 9040 pumps were functional, 5266 were non-functional and 544 needed repairs out of 14850 test data items. Based on this Tanzania Ministry of Water will be able to address the pumps accordingly.

1. **REFERENCES**

* <https://github.com/pancr9/Pump-It-Up/blob/master/Pump%20It%20Up%20-%20Data%20Cleaning.ipynb>
* <https://chrisalbon.com/machine_learning/model_evaluation/plot_the_validation_curve/>
* <https://www.kaggle.com/arthurtok/feature-ranking-rfe-random-forest-linear-models>